Intelligibility Enhancement of Speech in Noise

Intelligibility enhancement of synthetic speech in noise

Cássia Valentini Botinhão

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2013

Abstract

Speech technology can facilitate human-machine interaction and create new communication interfaces. Text-To-Speech (TTS) systems provide speech output for dialogue, notification and reading applications as well as personalized voices for people that have lost the use of their own. TTS systems are built to produce synthetic voices that should sound as natural, expressive and intelligible as possible and if necessary be similar to a particular speaker. Although naturalness is an important requirement, providing the correct information in adverse conditions can be crucial to certain applications. Speech that adapts or reacts to different listening conditions can in turn be more expressive and natural. In this work we focus on enhancing the intelligibility of TTS voices in additive noise. For that we adopt the statistical parametric paradigm for TTS in the shape of a hidden Markov model (HMM-) based speech synthesis system that allows for flexible enhancement strategies.

Little is known about which human speech production mechanisms actually increase intelligibility in noise and how the choice of mechanism relates to noise type, so we approached the problem from another perspective: using mathematical models for hearing speech in noise. To find which models are better at predicting intelligibility of TTS in noise we performed listening evaluations to collect subjective intelligibility scores which we then compared to the models' predictions. In these evaluations we observed that modifications performed on the spectral envelope of speech can increase intelligibility significantly, particularly if the strength of the modification depends on the noise and its level. We used these findings to inform the decision of which of the models to use when automatically modifying the spectral envelope of the speech according to the noise. We devised two methods, both involving cepstral coefficient modifications. The first was applied during extraction while training the acoustic models and the other when generating a voice using pre-trained TTS models. The latter has the advantage of being able to address fluctuating noise. To increase intelligibility of synthetic speech at generation time we proposed a method for Mel cepstral coefficient modification based on the glimpse proportion measure, the most promising of the models of speech intelligibility that we evaluated. An extensive series of listening experiments demonstrated that this method brings significant intelligibility gains to TTS voices while not requiring additional recordings of clear or Lombard speech. To further improve intelligibility we combined our method with noise-independent enhancement approaches based on the acoustics of highly intelligible speech. This combined solution was as effective for stationary noise as for the challenging com-

peting speaker scenario, obtaining up to 4dB of equivalent intensity gain. Finally, we proposed an extension to the speech enhancement paradigm to account for not only energetic masking of signals but also for linguistic confusability of words in sentences. We found that word level confusability, a challenging value to predict, can be used as an additional prior to increase intelligibility even for simple enhancement methods like energy reallocation between words. These findings motivate further research into solutions that can tackle the effect of energetic masking on the auditory system as well as on higher levels of processing.

Acknowledgements

I would like to thank the following people:

- My supervisors, Simon King and Junichi Yamagishi, for advice and guidance, for the time spent discussing new ideas, the technical help to implement them and for providing so many opportunities to improve and extend my work.
- The researchers from the LISTA project, in particular Martin Cooke, Yannis Stylianou, Elizabeth Godoy, Cassie Mayo and Yan Tang, for their helpful comments and for providing me with a framework of ideas, technical knowledge and data.
- Ranniery Maia for helping to structure the framework for cepstral manipulation and proof reading my draft dissertation report.
- Mirjam Wester for the support and supervision during the work that lead to the last chapter of this thesis and for proof reading this thesis.
- Adriana Stan for reading through this document and running a test viva.
- Vasilis Karaiskos for helping me set up my first few listening experiments scripts and for the time taking and recruiting participants.
- Heiga Zen for his suggestions and for providing the recordings of car noise.
- The SCALE project for providing the financial support and opportunities for professional training.
- Everyone from CSTR – and visitors – for providing such a friendly environment to work in and for their constructive criticism and suggestions in internal progress reports and practice presentations.
- My friends Simone Ferlin, Yunyun Ni and Marilia Maia for advice and encouragement.
- Alex Dawson, for proof reading, patience, food, advice and support during all these years.
- My family for supporting my decisions that lead me to be where I am and for the emotional support I needed to see it through to the end.

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213850 (SCALE).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Cássia Valentini Botinhão*)

Table of Contents

1 Introduction **1**

1.1 Organization of this thesis 2

2 Noise compensation strategies **6**

2.1 Hearing in noise 6

2.2 Speaking in noise 8

2.3 Speech modifications 10

3 HMM-based speech synthesis **12**

3.1 Vocoder 17

3.1.1 Source 17

3.1.2 Filter 20

3.1.3 Analysis 24

3.1.4 Synthesis 28

3.2 Acoustic model 30

3.2.1 Hidden Markov models 31

3.2.2 Training HMMs 32

3.2.3 Training synthesis models 34

3.2.4 Parameter generation 37

3.2.5 Adaptation 40

3.2.6 Oversmoothing 42

3.3 Subjective evaluation 43

3.3.1 Procedure 43

3.3.2 Types of listening test 44

3.3.3 Intelligibility studies 45

4 Evaluation of objective intelligibility measures **50**

4.1 Introduction 50

4.2 Objective intelligibiliy measures 54
4.2.1 Spectrum-based measures 55
4.2.2 Perceptually-motivated measures 57
4.2.3 Standards for quality and intelligibility 58
4.2.4 Perceptual model-based measures 59
4.3 Prediction of the intelligibility of modified speech in noise 60
4.4 Experiment I: synthetic speech and modifications based on the ideal binary mask 63
4.4.1 Experimental data 63
4.4.2 Speech modification 65
4.4.3 Listening set-up 66
4.4.4 Subjective intelligibility scores 66
4.4.5 Evaluation results 67
4.5 Experiment II: synthetic speech and modifications based on Lombard speech 70
4.5.1 Experimental data 70
4.5.2 Speech modifications 71
4.5.3 Listening set-up 74
4.5.4 Subjective intelligibility scores 75
4.5.5 Evaluation results 78
4.6 Discussion 80
4.7 Conclusion 84

5 Cepstral extraction using the glimpse proportion measure 85
5.1 Introduction 85
5.2 Maximum likelihood-based cepstral analysis 86
5.2.1 Unbiased estimator of the log spectrum 87
5.2.2 Cepstral coefficient extraction using UELS 88
5.3 The glimpse proportion measure 90
5.4 Proposed GP approximation 93
5.4.1 Evaluation of the proposed GP measure 95
5.5 GP-based cepstral coefficient extraction 95
5.5.1 Cost function 96
5.5.2 Steepest descent solution 96
5.5.3 Energy normalization 99

5.6 Evaluation . . . 103
5.6.1 Stimuli . . . 104
5.6.2 GP scores . . . 105
5.6.3 Acoustic analysis . . . 105
5.6.4 Listening experiment . . . 105
5.6.5 Results and discussion . . . 105
5.7 Conclusion . . . 106

6 Mel cepstral modification using the glimpse proportion measure 109
6.1 Introduction . . . 109
6.2 Mel cepstral coefficients . . . 111
6.3 GP-based Mel cepstral modification . . . 111
6.3.1 Cost function . . . 111
6.3.2 Steepest descent solution . . . 112
6.3.3 Energy normalization . . . 113
6.3.4 Distortion control . . . 115
6.4 Evaluation . . . 118
6.4.1 Voice building . . . 119
6.4.2 Convergence analysis . . . 121
6.4.3 Acoustic analysis . . . 122
6.4.4 Listening experiments . . . 130
6.4.5 Results and discussion . . . 132
6.5 Conclusions . . . 133

7 Evaluation of intelligibility enhancement methods 134
7.1 Stimuli material . . . 135
7.1.1 TTS . . . 135
7.1.2 Noise conditions . . . 136
7.1.3 Listening experiment . . . 136
7.2 Evaluation I: GP versus adaptation . . . 137
7.2.1 Acoustic analysis . . . 138
7.2.2 Listening experiment . . . 141
7.2.3 Results and discussion . . . 141
7.3 Evaluation II: Noise-dependent and -independent methods . . . 145
7.3.1 Methods . . . 145
7.3.2 Acoustic analysis . . . 147

7.3.3 Listening experiment 151
7.3.4 Results and discussion 151
7.4 Evaluation III: Adaptation and noise dependency 155
7.4.1 Methods 156
7.4.2 Acoustic analysis 156
7.4.3 Listening experiment 158
7.4.4 Results and discussion 158
7.5 Comparing across different listening tests 159
7.6 Conclusions 161

8 Using top-down information **163**
8.1 Spoken word recognition 164
8.2 Using word confusability to increase intelligibility 166
8.3 Measuring word confusability 167
8.3.1 Neighbourhood density 168
8.3.2 HMM-based distance 168
8.4 Stimuli material 170
8.5 Isolated word experiment 171
8.5.1 Word selection criteria 171
8.5.2 Word stimuli generation 172
8.5.3 Listening experiment design 172
8.5.4 Results 174
8.6 Sentence experiment: giver and receiver 179
8.6.1 Giver and receiver proposed classification 179
8.6.2 Sentence material 179
8.6.3 Modifications 181
8.6.4 Listening experiment design 183
8.6.5 Results 184
8.7 Sentence experiment: boosting one word 191
8.7.1 Sentence material 192
8.7.2 Modification 192
8.7.3 Listening experiment design 194
8.7.4 Results 195
8.8 Conclusion 200

9 Conclusions **202**

9.1 Contributions 203

9.1.1 Intelligibility prediction 203

9.1.2 Perception of synthetic speech in noise 204

9.1.3 Speech synthesis 205

9.2 Future work 206

A Objective measures of intelligibility **208**

B Spectral gains of Lombard speech **210**

C Hurricane Challenge results **211**

D Using top-down information **214**

Bibliography **220**

Chapter 1

Introduction

Speech is ultimately a communication tool, it gives us the possibility to interact with the world through the transmission of a message. Speech technology can improve communication by means of enhancing speech, making it more compact for transmission or storage and more robust to adverse conditions. In the field of speech technology, Text-To-Speech (TTS) systems provide the connection between text and speech, giving voices to those that have lost the ability to communicate and enabling machines to transmit messages through speech output. To maintain communication success, humans change the way they speak and hear according to many factors, like the age, gender, native language and social relationship between talker and listener. Other factors are dictated by how communication takes place, such as environmental factors like an active competing speaker or limitations on the communication channel. As in natural interaction, we expect to communicate with and use synthetic voices that can also adapt to different listening scenarios and keep the level of intelligibility high. Research in speech technology needs to account for this to change the way we transmit, store and artificially generate speech accordingly.

In this work, we focus on creating strategies to enhance intelligibility of synthetic speech in noise. In particular, we are interested in increasing intelligibility of Text-To-Speech voices generated by parametric statistical models (Zen et al., 2009). The statistical and parametric nature of hidden Markov model (HMM)-based speech synthesis offers a high degree of control over the generated speech. By modifying the models or extracted parameters we are able to control the acoustic characteristics of the generated speech without the need for new data. Intelligibility of HMM-generated synthetic speech is comparable to natural speech in clean environments (Yamagishi et al., 2008b), but in noisy environments the gap between natural and synthetic speech

grows with the level of the noise (King and Karaiskos, 2010).

One possible way of decreasing this gap is to mimic acoustic changes observed in highly intelligible speech. It is however not fully understood how these acoustic changes depend on the noise type and level. The naive solution of merely increasing the intensity of the speech signal is also not desirable as it can result in unpleasant and distorted speech as the noise level increases.

We hypothesize that it is possible to increase the intelligibility of speech in noise by modifying speech automatically according to the noise signal. We posit that this can be achieved using models of the human auditory system to provide estimates of the impact of noise on the processing of speech and to inform how speech should be enhanced.

1.1 Organization of this thesis

Speech produced in noise – Lombard speech – is more intelligible than speech produced in quiet due a mixture of conscious and reflex mechanisms. These mechanisms are in charge of controlling speech production according to what we hear and see. To permit the design of an automatic strategy it is important then to investigate what the mechanism of hearing speech in noise are as well as the possible strategies humans use to produce more intelligible speech. In **Chapter 2**, we present a survey of models of hearing and speech production in noise, pointing towards the models and strategies we will adopt in different parts of this thesis. We then describe existing methods for intelligibility enhancement, categorized into those that use additional recordings of highly intelligible speech, those that are based on acoustics and those that (similar to ours) modify speech automatically according to the noise signal.

Most of these methods and studies were performed with natural speech. To understand how synthetic speech is generated and how this might influence its intelligibility we present in **Chapter 3** the theoretical background supporting the TTS system used in this work: HMM-based speech synthesis. TTS systems convert text to speech through the completion of two tasks: text processing and waveform generation. Text processing is the conversion of the text message into relevant linguistic specifications that drive the process of waveform generation. The waveform generation is responsible for converting this specification in to an acoustic realization of speech. The acoustic model used for waveform generation in this work is based on HMMs. To train these models it is necessary to extract a linguistic specification and acoustic parameters from

speech corpora. The acoustic models are then trained to maximize the likelihood of the data and at generation time the linguistic specification is used to find the sequence of HMM models that best defines the utterance. The acoustic parameters extracted from this sequence are then used to construct a speech waveform. The intelligibility of HMM-generated synthetic speech is quite high in clean environments but similar to other types of synthesizers intelligibility in noise is highly compromised.

We hypothesize that it is possible to increase intelligibility automatically by using models of auditory processing but first we need to find which model to use. In **Chapter 4**, we describe two experiments that evaluate existing objective measures on the task of predicting intelligibility scores of HMM-generated speech in noise. With these experiments we want to discover an effective objective measure and a modification strategy. To choose the measure, we calculate the correlation between objective scores and subjective scores obtained in the listening tests and compare this correlation across many objective measures. To discover a promising strategy, we also evaluate the intelligibility of speech that has been modified. We chose to separately modify acoustic dimensions that are known to change in Lombard conditions: speaking rate, fundamental frequency, spectral sharpness and tilt. From these experiments we observed that a few measures – the ones based on an auditory model – perform quite well in the conditions used, including the case when speech was modified, with a correlation coefficient above 0.8. We also found that changes in the spectral envelope can be very effective across different noise types.

At this point, we have possible candidates for auditory-based measures and modification strategies. In **Chapter 5**, we present a new intelligibility enhancement method that uses one of these measures to extract cepstral coefficients that are used for training the synthesis models. We propose a cepstral extraction method based on the glimpse proportion (GP) measure. To use the GP measure for the task, we reformulate it into a measure that depends only upon the speech magnitude spectrum rather than the waveform. We then integrate this measure into an existing optimization method for cepstral extraction. Listening experiments with modified speech indicates that although GP values increase, subjective scores of intelligibility are not always higher. We hypothesized this happened because there was no proper control of how the glimpses were created.

To improve on these results we propose in **Chapter 6** a method for processing Mel cepstral coefficients – cepstral coefficients defined on the Mel scale. In this method, the modification happens at generation time from pre-trained synthesis models which

enables a solution for fluctuating noises like a competing speaker. In this reformulation, it is easier to control the amount of modification, or better yet limit the distortions introduced by modifying speech, and as a consequence improve subjective intelligibility. To do so, we limit the frequency resolution of the modification, modifying only the first few cepstral coefficients. Listening experiments with speech-shaped noise show that voices generated by modifying only the first two Mel cepstral coefficients are more intelligible and that this voice is as intelligible as voice built with Lombard speech data in the speech-shaped noise case. In a competing speaker scenario, intelligibility gains are smaller, even when Mel cepstral coefficients are obtained using the Lombard speech data.

In **Chapter 7**, we evaluate the GP-based method against other methods in a series of three large scale listening experiments with many participants and stimuli. The first evaluation compares the proposed modification with acoustic models adapted to Lombard speech recordings of that speaker. Results show that Lombard-adapted changes to duration and excitation signal can bring large intelligibility improvements in the competing speaker masker condition. The GP method obtains comparable gains in speech-shaped noise without requiring the additional recordings of speech produced in noise. In the second evaluation, we investigate whether it is possible to improve results in the competing speaker condition by combining the GP-based method with dynamic range compression (DRC), a strategy that reallocates energy across different time segments of speech to maximize intensity levels. Results indicate that adding DRC improves intelligibility in all noise scenarios. In the third experiment, we test a further combination: GP-based modifications applied to the spectral parameters, Lombard adaptation of excitation and duration parameters and finally the DRC applied to the synthesized waveform. With this combined strategy we improve intelligibility in the competing speaker scenario to match the performance in stationary noise. In both noises, for a medium SNR we increase intelligibility of speech by 4 dB of equivalent intensity gain.

We successfully created a strategy that increases subjective intelligibility of speech in noise averaged across words and listeners. There is however a great amount of variability in the intelligibility of words, which raises the question of whether all words should be treated in the same way, being driven only by the acoustics of the speech and the noise. In **Chapter 8**, we shift our focus to how to use top-down information for speech enhancement. One possible top-down source of information is the unit of the word. Word intelligibility depends on the listener's familiarity with the word and how

likely it is that this word is going to be said given the context (linguistic confusability). Word intelligibility also depends on how many phonetically similar words exist in the language or in the listener's lexicon (acoustic confusability). To avoid unnecessary modification a strategy to increase intelligibility should also account for these factors. In the experiments described in this chapter, we exploit word-level acoustic confusability to change the intensity of words in a sentence, constrained by a fixed SNR per sentence. We show that intelligibility of TTS in quiet of words in isolation can be quite low for words that have many neighbours (words that sound alike). In a sentence, intelligibility improves significantly across all words. We also show the potential for creating intelligibility enhancement strategies based on word-level information. This information is however still very difficult to predict as it involves semantic information and the acoustics of not only the word but also the other words in the sentence.

Chapter 2

Noise compensation strategies

We are interested in increasing the intelligibility of speech automatically according to the noise signal. For that, we need to find a method that can predict the effect noise has on intelligibility and find an effective strategy, i.e. which aspects of speech are worth enhancing. Before looking into evaluating different models and strategies we survey the literature on models of hearing and speaking in noise.

There are many perception studies which have investigated natural speech in noise, but rather than showing their findings we focus on the possible *mechanisms* that are involved in the process of hearing in noise, pointing to the ones we will further investigate in this thesis. The speech production in noise literature is also very extensive. We summarise what type of acoustic changes have been observed and in which listening circumstances. These findings will guide us in choosing intelligibility enhancement strategies and an interesting set of acoustic analyses that will help us evaluate them.

2.1 Hearing in noise

In a conversation, speakers and listeners adopt numerous strategies to compensate for additive noise. The listener for their part brings, consciously or not, extra effort to focus on the speaker over other sound sources. In many situations in fact, we are forced to deal with a great number of different sound sources and retrieve information from this "scene". The mechanism that explains why, and to a certain extent how, we are able to understand speech in a mixtures of sound sources concerns several different stages of processing. These mechanisms include: auditory grouping, glimpsing, linguistic adjustments and the regard for spatial and visual cues (Loizou, 2007).

Auditory grouping, also known as auditory streaming, is the capacity of a listener

to group together different time-frequency regions of speech and associate them to a single audio source (Bregman, 1990). This phenomenon can happen in two ways: simultaneous grouping, the capacity of grouping units appearing all at a certain time but at different spectral bands, and sequential grouping, which is the grouping of sound units that appear sequentially in time but at the same spectral band. In order to identify regions of speech that happen simultaneously as coming from the same source we can for instance track fundamental frequency (F_0) to identify regions that contain similar harmonics as coming from the same source. To group speech that appears sequentially it is important to track the evolution of speech features across time, in particular those features that are continuous and slowly changing such as formants and spectral peaks.

Another important auditory mechanism that aids source separation is so-called glimpsing, which is the ability to extract time-frequency regions where the corrupted speech signal is less masked and therefore less distorted (Cooke, 2003). This ability can help explain why stationary noises are stronger maskers than competing speakers, as the latter provide more gaps to the listener (Festen and Plomp, 1990). The glimpsing phenomenon also indicates that humans attend to changes in the local signal to noise ratio (SNR). It still remains unknown, however, what constitutes a useful glimpse and how to measure whether a certain region with high enough SNR is going to contribute to intelligibility gains. As we will see in Chapters 5 and 6, the glimpsing model will prove to be very useful for the sort of enhancement strategy we are focusing on. A more detailed explanation of how the model can be used to make a prediction of intelligibility will be described then.

When immersed in an adverse condition, listeners also make use of top-down processing: the linguistic information. This information can limit the number of possible guesses for what a word could have been given the lexicon and the context, which makes the decoding task much easier. Phone confusions in minimal word pairs that can arise due to noise can then be disambiguated with aid of the context information when we know that a certain word in that pair is more likely to appear in that context. A linguistic context advantage can appear in the shape of semantic and syntactic context as listeners expect sentences to be both semantically and syntactically correct. It is not clear how to use linguistic information to drive modification but a first attempt at using acoustic confusability – how confusable a word is given the lexicon of a language – will be presented in Chapter 8.

Other mechanisms that come into play when hearing speech in noise are based on the availability of spatial and visual cues. Spatial release of masking, that is when

target and masker sources are located in different regions of the acoustic space, can increase intelligibility by significant amounts due to the additional cues of intensity and time of arrival differences between ears, which can lead to an advantage of 2 to 7 dB equivalent intensity gain (Hawley et al., 2004). Visual information provided by the speaker can also increase recognition rates. Looking at the lips of the person talking can be crucial in distinguishing phones like /t/ and /p/ which are highly confusable in noise but are produced using different articulators. Such visual cues can give up to 11 dB in equivalent intensity gains (Macleod and Summerfield, 1987). In our work, we focus on creating strategies from the acoustics only, so we will not investigate a visual strategy. Spatial separation will also not be exploited as we assume no knowledge of the location of the noise source – although this could be estimated with the help of a microphone array.

2.2 Speaking in noise

To increase the success of communication, humans adapt to their immediate context by changing the way they produce speech. This adaptation can happen at different levels, that is, at an acoustic level with changes in phonation, place and manner of articulation or at a linguistic level, with changes in words and vocabulary (Lindblom, 1990; Picheny et al., 1985; Summers et al., 1988; Howell et al., 2006; Uther et al., 2007; Patel and Schell, 2008; Cooke and Lu, 2010).

The increase in vocal effort observed in speech produced in noise is generally called the Lombard effect (Lombard, 1911) and speech produced in noise is known as Lombard speech. Many acoustic changes have been reported for Lombard speech: an increase in intensity, increase in vowel duration, reduction in speaking rate, a shift in the energy distribution of the spectral content from low to middle and high frequency regions which results in flatter spectral tilt, increase in the first formant and in some studies increases in the second formant were also observed, increase in F_0 (both the average and the range) (Summers et al., 1988; Junqua, 1993; Hansen, 1996; Garnier et al., 2006; Lu and Cooke, 2008). It has also been reported that energy shifts from consonant to vowels (Junqua, 1993; Womack and Hansen, 1996; Garnier et al., 2006) and from semivowels to vowels and consonants (Hansen, 1996). More recently, Drugman and Dutoit (2010) showed that the glottal source is also significantly modified due to the increase in vocal effort: increases in F_0, decreases in the H1-H2 ratio (the ratio between the amplitude of the glottal spectrum at F_0 and at the second harmonic)

and increases in the amount of harmonics in the amplitude spectrum. Drugman and Dutoit (2010) also observed increases in the energy of specific spectrum frequency bands (which includes both the glottal and vocal tract changes): increases in E21 (the energy ratio between the frequency band 1-3 kHz and 0-1 kHz) and E31 (the energy ratio between band 3-8 kHz and 0-1 kHz); increases in E21 are substantially higher than E31. Also recently, Godoy and Stylianou (2012) presented a study on the loudness of Lombard speech, where it is observed that the loudness of Lombard speech is higher on average but not for all segments individually. Voiced segments are louder, unvoiced segments however are more quiet, in accordance with the energy shift from consonant to vowels observed in Junqua (1993).

Some studies have also observed that the Lombard effect changes according to many factors: the noise spectral content (Lu and Cooke, 2009a), the noise energetic and informational nature (Lu and Cooke, 2008; Cooke and Lu, 2010), the noise level (Summers et al., 1988; Lu and Cooke, 2008), the type of speaking task (whether it is read speech or conversational speech) (Aubanel et al., 2011), the presence of visual cues (Fitzpatrick et al., 2011) and the linguistic content (Patel and Schell, 2008).

Although noise dependencies have been observed, it is still not clear how the acoustic changes relate to the noise. Lu and Cooke (2009a) collected Lombard speech produced in low-pass and high-pass filtered noise and observed that, although flattening of the spectral tilt was presented in Lombard speech induced by both noise types, the flattening was not as strong in the presence of high frequency noise. Flattening the spectral tilt is a bad strategy to adopt in the presence of high frequency noise as it reallocates energy to the most masked region. These results adds to the discussion on how much of the Lombard effect is conscious and how well we can adapt to changes in the background noise given the physical constraints of our phonation system.

Lombard speech is known to be more intelligible than speech produced in quiet even when presented at the same level, that is at the same signal to noise ratio (Summers et al., 1988; Junqua, 1993; Lu and Cooke, 2008). It remains relatively unknown however which aspects of Lombard speech contribute to this intelligibility gain as well as how they relate to the characteristics of the environmental noise and the task involved. Although changes in F_0 have been observed, it is not yet clear what their role in the contribution to the intelligibility gain is. Lu and Cooke (2009b) noted that in the presence of speech-shaped noise, changes in F_0 do not contribute to intelligibility gains. To increase intensity, speakers involuntarily raise their vocal effort in two ways: by flattening the spectral slope (relative energy of upper harmonics increases)

and increasing F_0 itself. The increase in F_0 can then be seen as an involuntary (reflex) response to the increase in intensity.

2.3 Speech modifications

The different mechanisms of speech production in noise show that it is possible to modify speech in such a way that the mixture of speech and noise is more intelligible for the listener without an overall level increase. To emulate such an effect one could for instance modify speech produced in quiet by mimicking the acoustic changes seen in studies of speech produced in noise. Methods under this category include: boosting the consonant-vowel power ratio (an effect usually observed in clear speech) (Niederjohn and Grotelueschen, 1976; Skowronski and Harris, 2006; Yoo et al., 2007), spectral tilt flattening and formant enhancement (McLoughlin and Chance, 1997; Raitio et al., 2011a), manipulation of duration and prosody (Huang et al., 2010), increasing of duration, intensity and F_0 of content words (Patel et al., 2006) and both formant and loudness enhancement (Zorilă et al., 2012). Because it is not known to what extent the acoustic changes relate to the characteristics of the noise, these types of speech modifications are noise-independent.

Another strategy is to make direct use of available recordings of Lombard speech data through voice conversion techniques (Langner and Black, 2005) and adaptation techniques (Raitio et al., 2011a; Picart et al., 2013). This requires recordings of Lombard speech data from the speaker whose voice is to be synthesized. A different approach, described by Nicolao et al. (2012), also makes use of adaptation to map between normal, hypo, and hyper-articulated speech, without having to acquire more than plain speech data.

Recent work – much of which has been proposed through the course of this thesis – has also been carried out using estimates of the noise context for so-called noise-dependent methods. These approaches include modification of the local SNR (Sauert and Vary, 2006; Tang and Cooke, 2010), unit-selection unit cost based on the speech intelligibility index (SII) (Cerňak, 2006), optimisation of spectral power reallocation based on the SII (Sauert and Vary, 2010, 2011) and a global fixed optimization to maximize the glimpse proportion (GP) (Tang and Cooke, 2012) as well as different strategies for the insertion of small pauses (Tang and Cooke, 2011) and GP-based duration changes (Aubanel and Cooke, 2013). Recently, Taal et al. (2012) presented an optimisation algorithm based on a spectro-temporal perceptual distortion measure

and in Petkov et al. (2012) an algorithm based on a statistical model of speech was described. In Chapter 7, we present a series of evaluations comparing our strategies to some of the methods mentioned above, at which point we will present a more detailed description of them.

Providing speech that is easier to understand in noisy environments is of special interest in digital communication, where it is possible to modify the speech signal at both ends of the communication channel as long as the processing delay is low. During transmission degradation can arise from quantization and from the background acoustic noise of both the sending (far) or the receiving (near) end of the channel. Noise suppression algorithms can be applied to the far-end signal as a pre-processing stage to remove noise before transmission. It is also possible to improve the intelligibility of the transmitted quantized signal by applying post-processing techniques. In this scenario, speech intelligibility enhancement is often referred to as near-end or source-based speech enhancement, referring to where the enhancement process takes place. In this category we find a range of post-filtering techniques such as: formant structure and pitch peaks enhancement (Chen and Gersho, 1995; Grancharov et al., 2008), high-pass filtering (Hall and Flanagan, 2010; Jokinen et al., 2012), Lombard inspired modifications like spectral tilt flattening, formant sharpening and F_0 adaptive high-pass filtering (Jokinen et al., 2013), spectral power reallocation for maximizing the SII (Sauert and Vary, 2006, 2010) and spectral tilt modification followed by dynamic range compression (Erro et al., 2012). Due to the requirements of mobile communication these sort of techniques have to work with narrowband speech signals (speech signal sampled at 8 kHz), be robust to noise estimation and work at a low computational cost and low processing delay.

In synthetic speech intelligibility enhancement evaluation (Cooke et al., 2012, 2013) it is common to adopt a sentence-level energy constraint which will also be adopted in this thesis: the energy of the modified speech – calculated over the entire sentence – is normalized to be equal to the energy of the unprocessed speech signal. For speech transmission, however, evaluation is more strict: the energy per frame should not be modified. While one might argue against it, per sentence energy normalization allows for a wider range of long-term strategies to be applied such as boosting certain words, phonetic units or regions to the detriment of other parts of speech that can potentially be exploited in applications such as TTS and the reproduction of pre-recorded speech.

Chapter 3

HMM-based speech synthesis

In this chapter, we present the theoretical basis from which the contributions of our work, described in the chapters to follow, is structured upon. We describe in detail HMM-based speech synthesis in terms of the vocoder and the acoustic models used and then give an overview on how to evaluate synthetic speech.

Speech can be generated from text in a variety of ways. The first TTS methods proposed were constructed by rules on how speech sounds are produced. This was implemented by either following rules on the realization of acoustic components like the formants and the fundamental frequency (formant synthesizers), one such example is the DECtalk system proposed in (Klatt, 1980), or physical components like the position of the articulators (articulatory synthesizers). Rule-based TTS systems are created from prototypical rules of speech production that can create intelligible but very unnatural voices. The parametrization of production enables controllability, however devising rules for formant and articulator placement manually requires a great deal of expert knowledge. Instead of following production rules, the next generation of TTS systems create speech from the concatenation of natural speech components. These components are derived during the training of the system from a large database of several hours of speech. Concatenative systems were first proposed in the shape of fixed component units, diphone synthesizers (Moulines and Charpentier, 1990). A diphone is a segment defined from the middle of one phone to the middle of the subsequent phoneme. These segments were represented by linear predictive analysis components extracted during training (Moulines and Charpentier, 1990). As more storage and computing power became available, the second generation of concatenative systems appeared: unit selection systems (Sagisaka et al., 1992; Hunt and Black, 1996; Beutnagel et al., 1999). In unit selection, the segments – units – of concatenation

are variable in size. The best unit is selected according to the linguistic specification extracted from the text to minimize two costs: a unit cost – which segment best describes the text – and the join cost – which segment sequence generates the least join errors. Given a large enough speech database (at least 6 hours) unit selection voices' quality and naturalness can be quite high. Segmental quality is however compromised when unseen units are not represented correctly and segments are not correctly combined. The quality of the voice is strongly tied to the quality and coverage of the recordings used for building the voice which in turn limits the flexibility of the system. Controlling dimensions like speaker characteristics and speaking style require either a substantial amount of additional recordings or a great deal of signal processing. Issues with segmental quality will also affect the performance of any sort of post processing enhancement technique.

Proposed at the end of the 1990s, another paradigm for creating speech from text appeared based on units derived from statistical models, the statistical parametric TTS systems (Yoshimura et al., 1999; Zen et al., 2007a; Ling et al., 2006; Black, 2006). At synthesis time, the models are used to generate a low dimension parametric representation of speech. Instead of storing a large database of units this system represents units of speech by model parameters of lower dimensionality. The most widely used statistical model for statistical parametric TTS is the hidden Markov model (HMM), creating what is referred to the HMM-based speech synthesis systems. HMMs are used in other areas of speech technology like speech enhancement, conversion and quite extensively in the field of automatic speech recognition. Advances in this field led to many different methods and criteria for training, clustering and adapting HMMs, alongside freely available toolboxes such as HTK (Young et al., 2006) and HTS (Tokuda et al., 2009). Using statistical models as a choice for acoustic modelling has also influenced research in parametric representations of speech that can offer good interpolation and compression properties.

Due to its statistical and parametric nature, HMM-based speech synthesis presents many advantages over the other TTS paradigms:

– generalization: wider coverage of the acoustic space. Although still limited by the examples in the training data, it is able to generate waveforms that do not appear in this database through the combination of the extracted parameters;

– smaller footprint: storage of the statistics of acoustic models rather than waveform templates;

– versatility: new voices of different speakers and speaking styles can be easily obtained by transforming model parameters through well established model adaptation techniques that require small amounts of additional recordings;

– robustness: the quality of generated speech is more robust to variability in recording conditions and speaking quality as reported in Yamagishi et al. (2008a);

– unified learning: text and acoustic analysis can be jointly performed in an unified statistical approach;

– controllability: parametrization and context dependency allow for localized control strategies;

– multilingual support: a large recording database of a particular language is not required to build a voice with good quality.

The process of generating an acoustic realization from unseen texts can be divided into two main tasks: text analysis and waveform generation, the so-called front-end and back-end. The task of the front-end is to convert written text into a sequence of linguistic specifications. This feature extraction defines the linguistic aspect of speech that characterize its acoustic realization. Linguistic specifications can be for instance phonemes, syllables, pause and tone predictions. Systems that are used to obtain these linguistic specifications from text are: letter-to-phoneme conversion, phoneset assignment, part-of-speech tagger, phrase-break predictor and pitch-accent predictor. Although performed automatically through machine learning techniques, building each of these tools requires a degree of high level knowledge of the language as all this information is strongly tied to language and its regional variation. The linguistic specification can be represented in a series of *yes* and *no* answers obtained using these tools. The output of the front-end is a vector sequence of around 2000 binary values defining the linguistic specification of a phone. This information will be used to identify linguistic contexts that characterize acoustically similar segments of speech as represented by the acoustic models.

Given these linguistic specifications, the back-end is responsible for the generation of acoustic segments of speech. First attempts of back-ends as mentioned previously were constructed following speech production rules. With the increase of storage power and with the development of algorithms that can deal with larger amounts of data a second paradigm for back-end appeared: the data-base paradigm. First by concatenation of speech segments derived from a speech database – diphone and unit

selection synthesis – and secondly by generation of parametric sequences obtained by data-driven rule modelling – statistical parametric synthesizers.

HMM-speech synthesis systems generate speech by using HMMs for modelling vocoder parameters (Zen et al., 2009). The models are trained with parameters extracted from natural speech, to maximize the likelihood of the training data. The source can be represented by the fundamental frequency and the aperiodicity band energies and the spectral envelope by Mel generalized cepstral coefficients (Tokuda et al., 1994) or line spectral pairs (Itakura, 1975a). The block diagram structure of an HMM-based speech synthesis system is displayed in Fig. 3.1. In the training part at the *analysis* stage, a set of parameters is extracted from the natural speech database to form the observation vector $\boldsymbol{O}$. These parameters describe the excitation signal and spectral envelope separately. HMM models are then used to model these observation vectors. The parameter set characterising the model λ is obtained by maximizing the likelihood of the training data:

$$\lambda_{max} = \arg\max_{\lambda} P(\boldsymbol{O}|\lambda, \mathcal{W}) \tag{3.1}$$

where $\boldsymbol{O}$ is the set of observation vectors – representing the acoustic parameters extracted by the vocoder – and $\mathcal{W}$ is the linguistic specification sequence corresponding to $\boldsymbol{O}$.

The system is trained with linguistic and prosodic contexts contained in the labels. Each distinct linguistic specification would ideally be represented by a separate model, but such context-dependency would result in a vast number of possible models that could not be covered by a database except in a sparse manner. To improve context coverage and allow for unseen context, it is necessary to decrease the possible number of models by means of iewstate clustering. In this thesis, clustering is performed with decision trees, where splits are made using the linguistic specification so that the log-likelihood of the data given the tied model of the tree leaves is maximized. First, monophone HMM models that are context independent are trained, then context dependent HMMs are trained. Their states are then clustered. The parameters of the clustered states are then further trained.

The parameter generation (Tokuda et al., 2000) is achieved by the maximization of the output probabilities given a trained model:

$$\boldsymbol{O}_{max} = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|w, \lambda_{max}) \tag{3.2}$$

The generated observation sequence is fed to the *synthesis* part of the vocoder responsible for reconstructing the waveform from this low dimensional representation

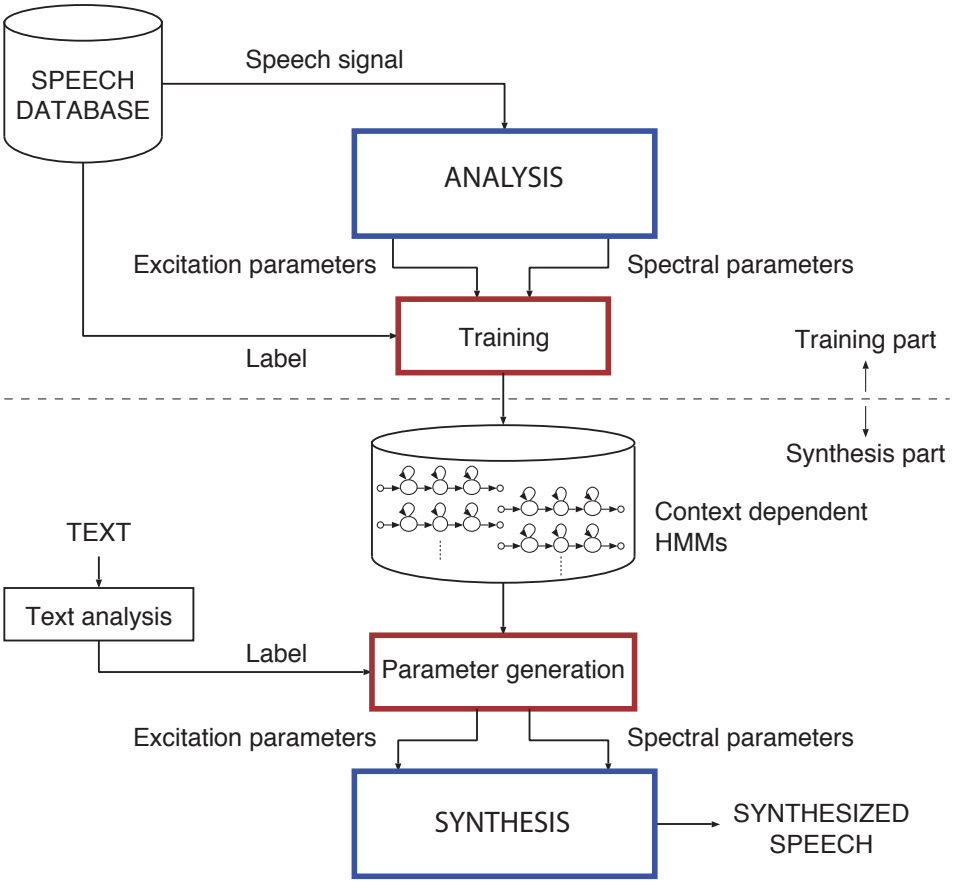


Figure 3.1: Block diagram of an HMM-based speech synthesis adapted from Zen et al. (2009), vocoder components in blue and acoustic model related components in red.

of the spectral envelope and the excitation signal.

In the next two sections, we will describe this process in more detail, first the vocoder analysis and synthesis methods and then the acoustic model and the operations of training, generating and adapting. We will focus here on the methods that composed the system used in this thesis, pointing to the literature regarding possible alternatives.

3.1 Vocoder

The vocoder is the mechanism responsible for analysing and synthesizing speech through the use of a intermediate representation of the speech waveform. The most common vocoder used in statistical parametric TTS is based on the source-filter model of speech production. This model assumes that speech is generated by passing a signal, the source, referred to as the excitation signal, through a filter that represents the vocal tract, as presented in Fig. 3.2. It is desirable to find parametrizations that provide a high degree of separation between source and filter so that independence assumptions are better met. There are however other vocoders based on a sinusoidal model for speech (Hemptinne, 2006; Banos et al., 2008), for example. In this section, we present details on how the source-filter model can be designed in terms of the choices for what represents the source and the filter, pointing to the ones we use in this thesis, following a more detailed presentation of the analysis and synthesis of the vocoder we used.

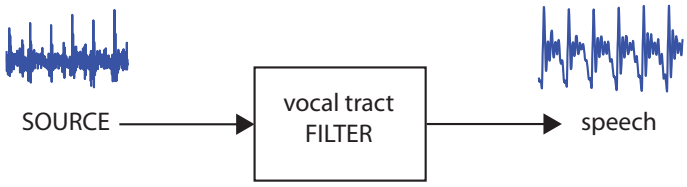


Figure 3.2: Source filter model of speech production.

3.1.1 Source

The source is represented by the excitation signal, i.e. the signal that excites the vocal tract filter. The ideal excitation signal is the residual signal defined according to the vocal tract filter by inverse filtering as shown in Fig. 3.3. The inverse filtering operation converts speech into the residual signal by applying the inverse of the filter that models the vocal tract given an available speech waveform. Any residual error that arises from representing speech by using a minimum-phase filter (where the stability of the inverse operation is guaranteed) will be contained in the residual. Perfect reconstruction can be achieved if the residual signal is not transformed: speech is synthesized by filtering the residual signal using the vocal tract filter. The residual signal can be better compressed than the speech waveform as it contains less information. This property has motivated a series of speech coding techniques – techniques for representing speech at a lower rate – based on residual coding. Instead of transmitting or storing the speech waveform, parameters of the vocal tract filter and a lower dimensional representation

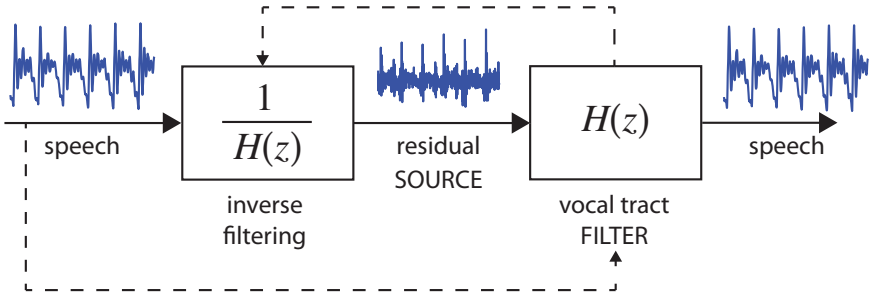


Figure 3.3: Residual signal. The vocal tract filter is calculated from speech and its inverse is used for obtaining the residual signal.

of the residual signal are obtained, leading to techniques like code-book excited linear prediction (CELP) (Schroeder and Atal, 1985).

The ideal excitation signal as said previously is the residual signal obtained from the speech waveform via inverse filtering. It is not straightforward how one could use residual signals for statistical parametric TTS as the synthesis models do not provide the representation of the speech waveform. An attempt towards using residual signals for HMM-based synthesis is described in Maia et al. (2007). The authors propose the extraction of filters that, at generation time, would be used to create a mixed excitation signal. These filters are obtained by maximizing the likelihood of the residual signal obtained from the training set, so that the excitation signal generated by these filters and the residual signal extracted from inverse filtering are similar. In that way, the excitation signal is obtained by minimizing the residual error of the source filter model. Another example of using the residual signal is described in Drugman et al. (2009) where the residual signals obtained at the training stage are stored in a codebook. During generation this codebook is accessed by a lower dimensional representation of the residual signal generated from the synthesis models.

The alternative way to represent the source is to use a parametric representation of the excitation signal. The most basic parametrization creates the so called simple excitation signal, seen in Fig. 3.4, where the source is modelled as either a pulse train (harmonics) or white noise (stochastic), representing the periodic and aperiodic segments of speech. This was the excitation signal used in the first HMM-based statistical parametric TTS systems (Yoshimura et al., 1999). This is a very simplified way of modelling the source as it is known that many speech segments contain both periodic and aperiodic excitation components, creating low quality buzzy vocoded speech. Other parametrizations have been proposed based on a mixed signal: excitation is created by a mixture of periodic and aperiodic signals as presented in Fig. 3.5. This type

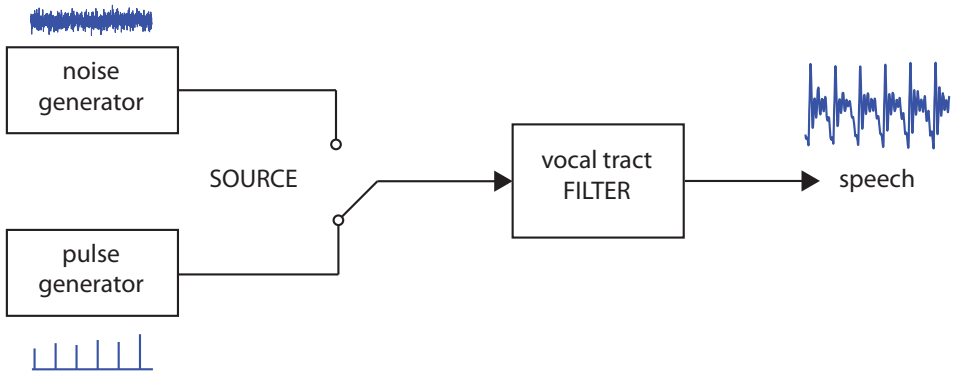


Figure 3.4: Simple excitation: the source is modelled by either a pulse train (voiced) or a random noise (unvoiced) sequence.

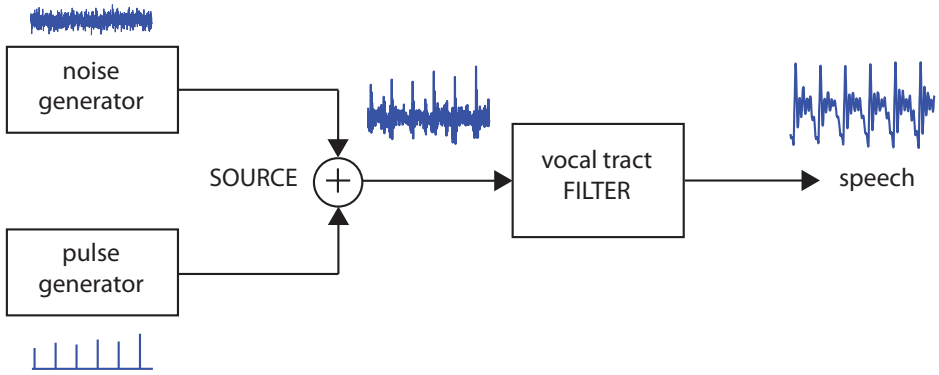


Figure 3.5: Mixed excitation: the source is modelled as a mixture of pulse and noise.

of excitation signal was first proposed for statistical parametric TTS in Yoshimura et al. (2001) and since then has been widely adopted. It is a design decision how to create these signals and how to mix them. The vocoder used in our experiments is based on a multiband mixed excitation signal where the excitation is represented by a spectral weighted mixture of a pulse and of white noise.

Yet another way of representing the source is to characterize the signal generated by the vocal folds: the glottal signal. Two notable methods of using glottal source signals for HMM-based speech synthesis are based on a library of glottal pulses (Raitio et al., 2008, 2011b) and a parametric model of the glottal-flow derivative (Cabral et al., 2007, 2008).

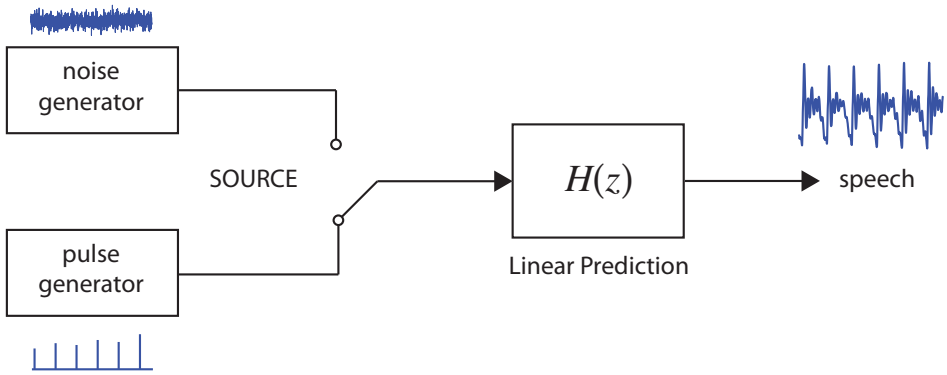


Figure 3.6: Linear predictive analysis.

3.1.2 Filter

3.1.2.1 Linear predictive analysis

Linear predictive analysis defines the spectral envelope $H(z)$ as an all-pole filter:

$$H(z) = \frac{G}{A_M(z)} = \frac{G}{1 - \sum_{m=1}^{M} a_m z^{-m}} \tag{3.3}$$

where $\{a_i\}_{i=1}^{M}$ are the linear prediction (LP) coefficients and G is a gain factor. Another way of interpreting this in the context of the source-filter model is that speech can be represented by an autoregressive (AR) process (Itakura and Saito, 1970). A process is an AR process of order M when its current value is a linear combination of its past M values plus an innovation component represented by white noise:

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{m=1}^{M} a_m z^{-m}} \tag{3.4}$$

$$s(n) = \sum_{m=1}^{M} a_m s(n-m) + Ge(n) \tag{3.5}$$

where $S(z)$ and $E(z)$ are the $\mathcal{Z}$-transforms of the speech signal $s(n)$ and the excitation signal $e(n)$.

To calculate the LP coefficients one can use the autocorrelation method. Given LP coefficients and a source, speech can be vocoded through a filtering operation, as presented in Fig. 3.6. Stability of the synthesis filter $H(z)$ is guaranteed when the LP coefficients are obtained via the autocorrelation method, but the process of statistical modelling, that can be seen as a process of averaging, does not guarantee that the generated coefficients will create stable synthesis filters. To avoid this problem, one can represent the spectral envelope in the linear prediction paradigm by modelling an

intermediate representation where stability can be guaranteed more easily. One such intermediate representation is the set of so called line spectral pairs (LSP) (Itakura, 1975a).

The motivation behind the LSP parametrization is that the vocal tract, as modelled by an AR process, can be represented by two higher order polynomials $P(z)$ and $Q(z)$ that correspond respectively to the resonance conditions of the vocal cavity: closed and open glottis (Deller Jr. et al., 2000). The spectrum of the linear prediction coefficients can be represented by a symmetric and a non symmetric part:

$$A_M(z) = \frac{P(z)+Q(z)}{2} \tag{3.6}$$

where $P(z)$ and $Q(z)$ are the M+1 order polynomials named the antisymmetric and symmetric, referring to the complete closure and the complete opening of the vocal folds, represented by the additional filter coefficient.

The line spectral pairs or frequencies (LSP or LSF) are the roots of the $P(z)$ and $Q(z)$ polynomials, that are so defined:

$$P(z) = A_M(z) + z^{-(M+1)}A_M(z^{-1}) \tag{3.7}$$

$$Q(z) = A_M(z) - z^{-(M+1)}A_M(z^{-1}) \tag{3.8}$$

The roots of the polynomial are interlaced with each other on the unit circle so to find the LSP it is sufficient to find their angles.

A line pair defines a resonator, a peak or a valley in the spectral envelope. The distribution of the pairs across the frequency domain relates to the distribution of the energy of the spectral envelope: frequency bands densely populated with LSPs are more energetic. LSP have good quantization and interpolation qualities (Koishida, 1998). If the roots of the polynomials lie on the unit circle and are interlaced than the roots of $A(z)$ are inside the unit circle which ensures stability of the synthesis filter. For speech synthesis, the generation of LSPs that are not interlaced might lead to synthesis filters that are not stable, degrading the quality of synthesized speech. Although stability checks in the LSP domain are much easier to perform than in the LP domain, in other filter representations like the minimum-phase cepstral coefficient stability is always guaranteed.

3.1.2.2 Cepstrum analysis

According to the source-filter model, the speech signal $s(n)$ is the output of a linear filter – that has $h(n)$ as the impulse response – whose input is the excitation signal

$e(n)$:

$$s(n) = e(n) * h(n) \tag{3.9}$$

The Fourier transform $\mathcal{F}\{.\}$ turns the convolution operation $*$ into a multiplication operation and the logarithm in the spectral domain further turns it into a summation:

$$\mathcal{F}\{s(n)\} = \mathcal{F}\{e(n)\}\mathcal{F}\{h(n)\} \tag{3.10}$$

$$S(e^{j\omega}) = E(e^{j\omega})H(e^{j\omega}) \tag{3.11}$$

$$\log S(e^{j\omega}) = \log E(e^{j\omega}) + \log H(e^{j\omega}) \tag{3.12}$$

The complex cepstrum $c(m)$ of $s(n)$ is calculated as the inverse Fourier transform of its log spectrum:

$$c(m) = \mathcal{F}^{-1}\{\log S(e^{j\omega})\} \tag{3.13}$$

$$= \mathcal{F}^{-1}\{\log E(e^{j\omega})\} + \mathcal{F}^{-1}\{\log H(e^{j\omega})\} \tag{3.14}$$

$$= c_e(m) + c_h(m) \tag{3.15}$$

where $c_e(m)$ and $c_h(m)$ are the complex cepstrum of $e(n)$ and $h(n)$.

The complex cepstrum is an infinite and non-causal sequence. If $s(n)$ is a real sequence then its complex cepstrum is also a real sequence. The term complex refers to the fact that the complex cepstrum is calculated from the inverse Fourier transform of the logarithm of the spectrum, a complex sequence. The real cepstrum $\hat{c}(m)$, often simply referred to as the cepstrum, is calculated from the log magnitude spectrum instead:

$$\hat{c}(m) = \mathcal{F}^{-1}\{\log |S(e^{j\omega})|\} \tag{3.16}$$

As the real cepstrum does not contain any phase information of the signal it is not possible to reconstruct a signal from it. The real cepstrum is attractive as it is a causal sequence which guarantees a causal synthesis filter. The real cepstrum can be obtained from the complex cepstrum as follows:

$$\hat{c}(m) = \frac{c(m) + c(-m)}{2} \tag{3.17}$$

If the $s(n)$ is considered to be a minimum-phase sequence, its complex cepstrum becomes a causal sequence and it is possible to calculate it from the real cepstrum as follows:

$$c(m) = \begin{cases} 2\hat{c}(m) & \text{if } m > 0 \\ \hat{c}(m) & \text{if } m = 0 \\ 0 & \text{if } m < 0 \end{cases}$$

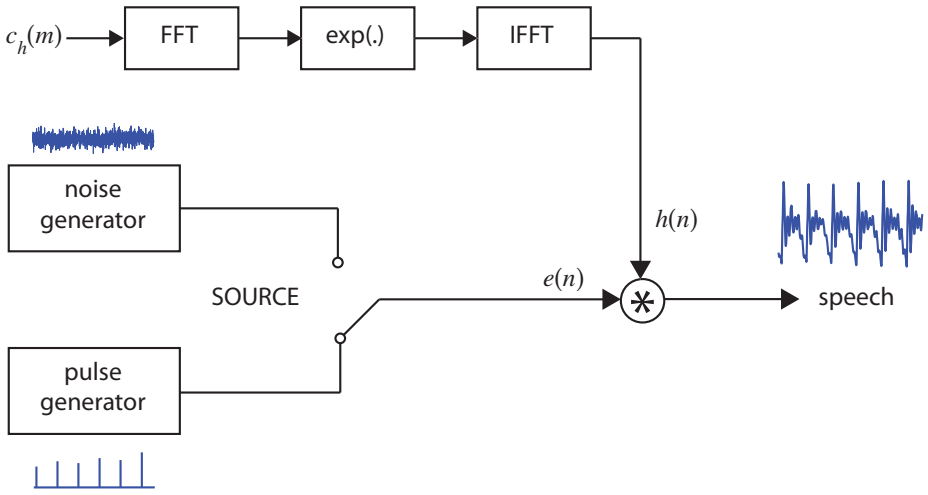


Figure 3.7: Cepstrum analysis.

A cepstrum calculated using the minimum-phase assumption is called the minimum-phase cepstrum, which is the cepstral representation that we use in this thesis. However the minimum-phase assumption – a synthesis filter whose poles and zeros are all inside the unit circle and is therefore, together with its inverse, both causal and stable – is not strictly necessary for TTS as causality is not a strong requirement. Recently, Maia et al. (2013) proposed how one could use the complex cepstrum for statistical parametric speech synthesis by incorporating the additional phase information from the decomposed all-pass filter into the excitation signal.

The log-magnitude spectrum of a signal can be represented by the summation of cosines weighted by the complex cepstrum:

$$\log|S(e^{j\omega})| = \sum_{m=0}^{\infty} c(m)\cos(m\omega) \tag{3.18}$$

The cepstral domain can then be seen as a frequency representation of the log-magnitude spectrum: high frequency fluctuations of the spectrum are represented by the higher order cepstral coefficients, while the low frequency fluctuations are represented by low order coefficients.

This is an especially attractive property of the cepstrum for modelling speech signals as the spectral envelope is characterized by low frequency resolution fluctuations while the spectrum of the excitation signal presents high fluctuation. These distinctive fluctuation patterns implies that $c_e(m)$ and $c_h(m)$ are not covering the same cepstrum region and can therefore be separated by a filter (liftering) operation. As presented in Fig. 3.7, from the estimated value of $c_h(m)$ it is possible to obtain the spectral envelope

and its impulse response:

$$H(e^{j\omega}) = \exp \mathcal{F}\{c_h(m)\} \tag{3.19}$$

$$h(m) = \mathcal{F}^{-1}\{\exp \mathcal{F}\{c_h(m)\}\} \tag{3.20}$$

When pitch-synchronous analysis is performed with an analysis window set to two pitch periods, the excitation signal can be considered to be a unit impulse response, implying $c_h(m) \approx c(m)$. The short-term spectral envelope can be estimated from a truncated M-order $c(m)$ sequence as follows:

$$H(e^{jw}) = \exp \sum_{m=0}^{M} c(m)e^{-jmw} \tag{3.21}$$

3.1.3 Analysis

During analysis, the parameters that represent the excitation signal and the spectral envelope are extracted from the speech waveform. These parameters are extracted separately as presented in the diagram of Fig. 3.8 by using the high quality vocoder known as STRAIGHT (Kawahara et al., 1999). STRAIGHT extracts a high dimensional representation of the spectrum and the aperiodicity contained in the speech signal which is then converted to a lower dimension representation that is easier to model. In the following sections we will describe each of these blocks in more detail.

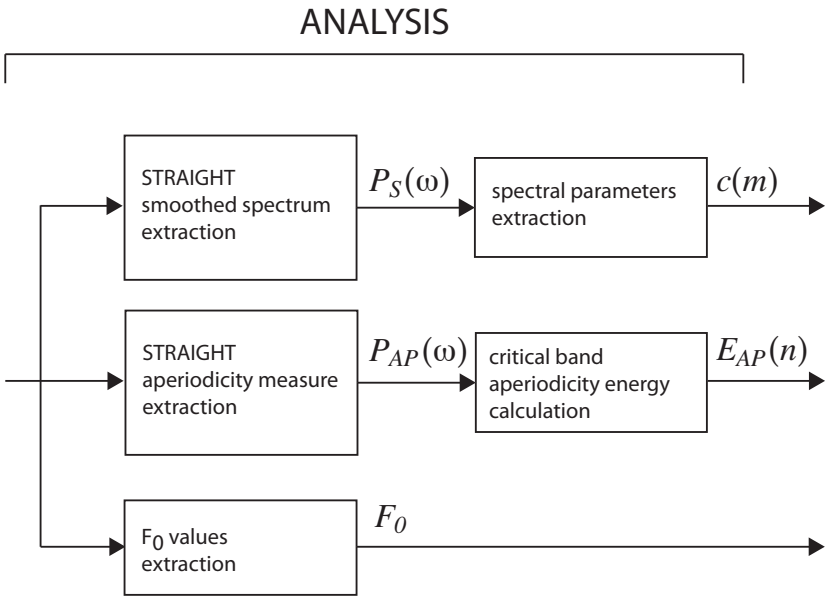


Figure 3.8: The analysis structure of the vocoder used in this thesis.

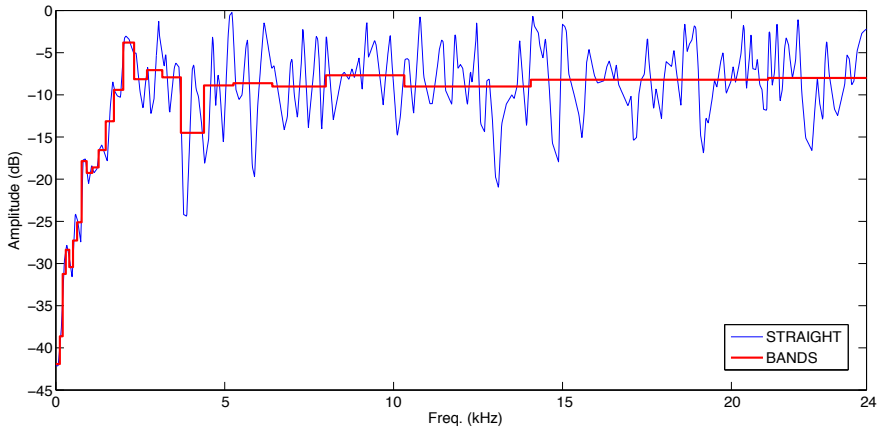


Figure 3.9: Aperiodicity spectrum of order 2049 calculated by STRAIGHT for a voiced speech frame and 25 energy bands linearly spaced on the Bark scale, referred to as the aperiodicity band energy $E_{AP}(n)$.

3.1.3.1 Excitation signal

The parameters that describe the excitation signal in the system are the aperiodicity band energy $E_{AP}(n)$ and the fundamental frequency F_0. If the excitation signal is represented by simple excitation then extraction and modelling of F_0 is sufficient. In this work, we use a multiband mixed excitation signal to represent the source signal, that is, the excitation signal is represented by a weighted mixture of periodic and aperiodic components. The amount of voicing, that is the weight given to the periodic signal, is controlled by the aperiodicity band energy parameters. These energy values are calculated from the higher dimensional aperiodicity measure $P_{AP}(\omega)$ extracted by STRAIGHT (Kawahara et al., 2001). STRAIGHT extracts this measure from $|S(\omega)|^2$, a power spectrum representation that contains the harmonic structure. Two smoothed spectral envelopes: the upper envelope S_U, constructed by connecting the peaks of the spectrum, and the lower envelope S_L, constructed by connecting the valleys, are calculated from the power spectrum representation. The aperiodicity measure is averaged across the spectral frequencies of the ratio between the lower and upper spectral representations and is weighted by the power spectrum $|S(\omega)|^2$:

$$P_{AP}(\omega') = \frac{\int \omega_{ERB}(\omega;\omega')\ |S(\omega)|^2\ \Gamma\left(\frac{|S_L(\omega)|^2}{|S_U(\omega)|^2}\right) d\omega}{\int \omega_{ERB}(\omega;\omega')\ |S(\omega)|^2 d\omega} \tag{3.22}$$

where $\omega_{ERB}(\omega;\omega')$ refers to a simplified auditory filter that selects a frequency band around ω' from the integral thus smoothing the power spectrum at centre frequency ω'. The symbol $\Gamma(.)$ refers to a table-look-up operation built from known aperiodic

signals.

If the ratio between the lower and the upper envelope is high, i.e. the upper envelope is just slightly above the lower envelope, then the value of the aperiodicity measure is high (Kawahara et al., 2001).

Aperiodicity band energies referred to as $E_{AP}(n)$ in Fig. 3.8 are calculated by taking the energy of the aperiodicity measure contained in a set of frequency bands indexed by n. Fig. 3.9 presents the aperiodicity spectrum (aperiodicity measure) of order 2049 and the 25 aperiodicity band energies calculated for a voiced segment. The bands are spaced linearly on the Bark scale.

To create the periodic component of the excitation signal the fundamental frequency F_0 needs to be extracted. In STRAIGHT, F_0 is extracted by the the TEMPO algorithm proposed in Kawahara (1997), which takes the fundamental frequency as the central frequency of the Gabor filter that presents the highest signal-to-noise ratio. However, in this work, we use the ESPS – SNACK toolkit – implementation based on the Robust Algorithm for Pitch Tracking (RAPT) (Talkin, 1995) to calculate F_0 instead. The RAPT algorithm is based on a normalized cross correlation function (NCCF). The RAPT algorithm works in two passes. First it processes the framed signal at a lower sample frequency rate and locates the peaks of the NCCF calculated over the whole range of interest. Then the NCCF of the higher sampling frequency signal is calculated around the neighbours of these collected peaks. This provides a better resolution whilst still keeping a lower computational complexity. For this method, a dynamic programming postprocessing step is performed to select between pitch candidates across time.

3.1.3.2 Spectral envelope

To calculate the spectral envelope parametrization, first the speech spectrum is extracted using STRAIGHT (Kawahara et al., 1999). Kawahara et al. (1999) claim that STRAIGHT can extract a spectrum representation that is less affected by the periodicity contained in the signal, compared to for instance the short term Fourier analysis-derived spectrum. The STRAIGHT-derived spectrum provides a spectral envelope representation with maximized separation from the fundamental frequency, making the independence between the spectral and excitation streams stronger.

To calculate the smoothed spectrum, STRAIGHT uses a time frequency analysis based on two window functions that are both pitch-adaptive – the length is set to two times the pitch period. The shape of the first window is set to be the convolution of

a Gaussian function and a second order cardinal B-spline function, offering similar resolutions in the time and frequency domains. The length of this window is set to be twice the pitch period in order to minimize the effect of the pitch period on the short term analysis, which makes frequency smoothing more robust to errors in estimating F_0. While the first window attenuates the periodic interference by smoothing the peaks of the spectrum the second window smooths the valley areas. The second window is constructed from the first window and it acts as a compensation window resolving the spectrum holes that the first window creates. The window is computed by multiplying the first window by a sinusoid function that produces maximas where the original spectogram has holes. The smoothed power spectrum is obtained by the weighted squared summation of the power spectra obtained using the original window $P_O^2(\omega)$ and the compensation window $P_C^2(\omega)$ where the weight ξ is set to minimize the temporal variation of the resulting smoothed spectrum:

$$P_S(\omega) = \sqrt{P_O^2(\omega) + \xi P_C^2(\omega)} \tag{3.23}$$

The STRAIGHT power spectrum is a high dimensional representation and for statistical modelling a lower dimensional representation is better. To extract a lower dimensional representation of the smooth spectrum one can use representations like linear prediction coefficients or cepstral coefficients, described in the Section 3.1.2. This lower dimensional representation is referred in Fig. 3.8 as the spectral parameters $c(m)$. The unifying approach of Mel Generalized Cepstral (MGC) analysis (Tokuda et al., 1994) incorporates linear prediction, cepstral analysis and frequency warping, all in one equation:

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_\alpha^{-m}\right)^{1/\gamma} & , 0 < |\gamma| \leq 1 \\ \exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m) z_\alpha^{-m} & , \gamma = 0 \end{cases} \tag{3.24}$$

where $c_{\alpha,\gamma}$ are referred to as the MGC coefficients and the all-pass frequency warping operation is defined as:

$$z_\alpha^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{3.25}$$

The parameter α controls the frequency warping. If α is equal to zero then the frequency axis is linear, if not, some sort of transformation in that domain is occurring. The parameter γ controls the logarithmic function, if $\gamma = 1$ then the function is linear and no transform is applied. In this case, the spectrum is modelled as an all-zero model. If $\gamma = -1$, the spectrum is modelled as an all-pole function using linear prediction

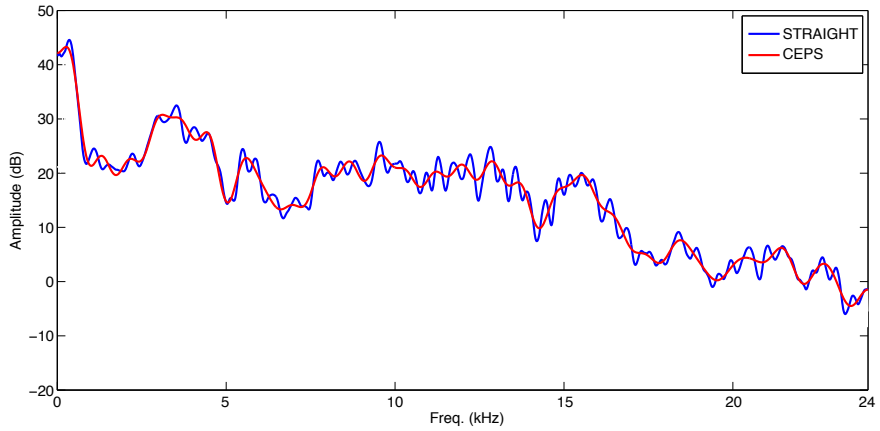


Figure 3.10: STRAIGHT spectrum of order 2049 and spectrum calculated from cepstral coefficients of order 59 extracted from the STRAIGHT spectrum.

(LP) analysis and if $\gamma = 0$ the transform applied to the spectrum is logarithmic and the spectrum is modelled through cepstral coefficients (cepstral analysis).

To extract MGC coefficients $c_{\alpha,\gamma}(m)$ from the smoothed spectrum $P_S(\omega)$, a method based on the unbiased estimator for the log spectrum (UELS) by Imai and Furuichi (1988) is used. In Chapter 5, we show how to extract cepstral coefficients using this method (Tokuda et al., 1995). The UELS-based extraction method has been extended to other parameters: Mel cepstral coefficients (Fukada et al., 1992), generalized cepstral coefficients (Tokuda et al., 1989) and Mel generalized cepstral coefficients (Tokuda et al., 1994). Fig. 3.10 presents the spectrum extracted using STRAIGHT and the spectrum calculated with cepstral coefficients extracted using the UELS method from the STRAIGHT spectrum.

From the generalized spectral envelope coefficients ($\gamma = 0$) it is possible to derive the Mel generalized LSP (MGC-LSP). MGC-LSP are the roots of the polynomials defining the filter obtained when using the generalized form of the spectral envelope. The process of obtaining the so called MGC-LSP is: extraction of MGC then conversion from MGC to MGC-LSP; a full derivation can be found in Koishida (1998).

3.1.4 Synthesis

The synthesis component of a vocoder is responsible for reconstructing the speech waveform from the intermediate model representation. In the system that we use this entails two operations: spectral envelope calculation and excitation signal creation. The speech waveform is obtained by the inverse Fourier transform of the multiplication

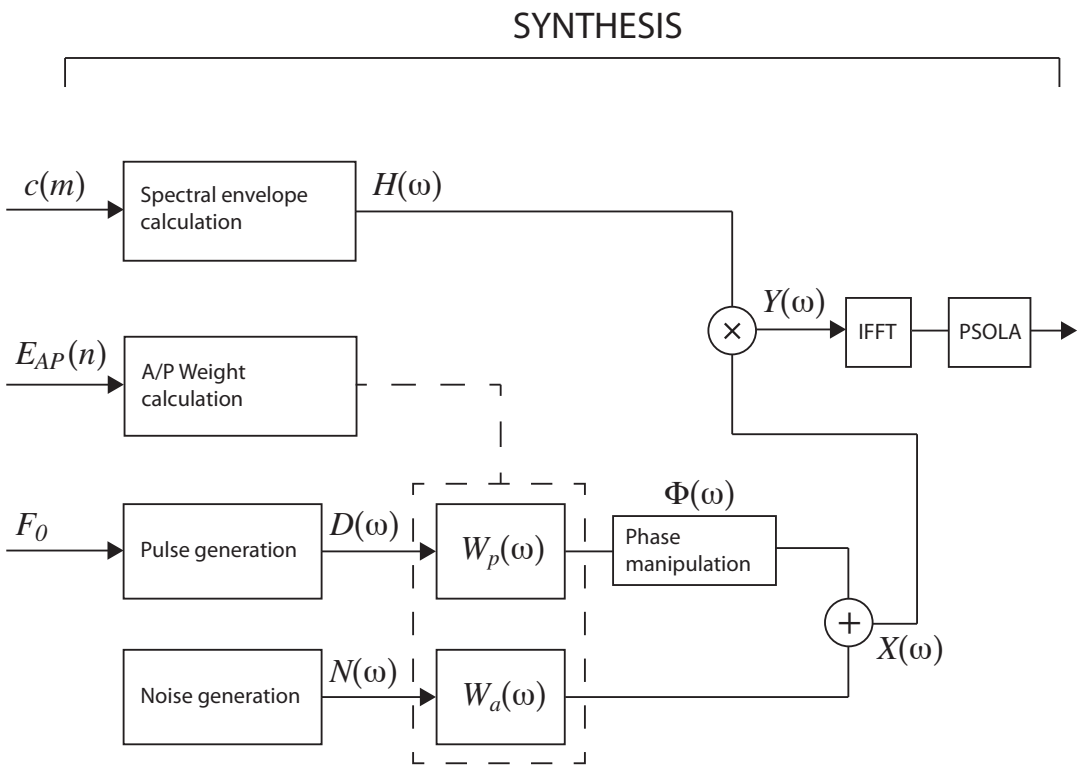


Figure 3.11: The synthesis structure of the vocoder used in this thesis.

of these two frequency representations. The process is pitch synchronous and the last operation is the overlap and addition of the analysis windows to the construction of the speech waveform.

Fig. 3.11 shows the synthesis procedure using the multiband mixed excitation signal. Although not explicitly shown in the block diagram, this process is pitch-synchronous, i.e. the window length is equal to twice the length of the pitch period for voiced segments and a fixed length for unvoiced ones. Fig. 3.11 shows the synthesis of a voiced segment, where the excitation signal is composed of the mixture of aperiodic (noise) and periodic (pulse) frequency representations. For unvoiced sounds the excitation is modelled as white noise only.

To obtain the excitation signal of voiced segments, the Fourier transform of the periodic and the stochastic sources are each multiplied by stepwise functions W_p and W_a respectively. This stepwise weighting function is obtained from the aperiodicity band energies in the following way (Yamato et al., 2006):

$$W_a(\omega) = \min(1, \mathcal{T}(E_{AP}(\omega)) \tag{3.26}$$

$$W_p(\omega) = \sqrt{1 - W_a^2(\omega)} \tag{3.27}$$

where $\mathcal{T}(.) = \frac{\mathcal{L}(.) - \mathcal{L}(0)}{\mathcal{L}(1) - \mathcal{L}(0)}$ and $\mathcal{L}(.)$ is a sigmoid function.

To attenuate the buzziness introduced by using a pulse as the periodic signal, the phase of its weighted frequency representation is adjusted by an all-pass filtering operation presented in Fig. 3.11 as the phase manipulation block with frequency response $\Phi(\omega)$. This phase adjustment introduces signal dispersion through group delay manipulation, turning the pure pulse into a shape that better resembles the glottal pulse. Details of how this filter is constructed can be found in Kawahara (1997).

Speech is synthesized in the frequency domain $Y(\omega)$ as follows:

$$Y(\omega) = X(\omega)H(\omega) \tag{3.28}$$

$$X(\omega) = \sqrt{1/F_0}\, D(\omega)\Phi(\omega)W_p(\omega) + N(\omega)W_a(\omega) \tag{3.29}$$

where $D(\omega)$ is the Fourier transform of the delta pulse, $N(\omega)$ is the discrete Fourier transform of white noise. The spectrum $H(\omega)$ is calculated from the spectral parameters $c(m)$ using the discrete frequency version of Eq.(3.24).

The noise is modelled by a random sequence with zero mean and unit variance. For the impulse train to have the same energy as the noise signal, the pulse is multiplied by $\sqrt{1/F_0}$.

Alternatively, speech can be synthesized in the time domain by a filter operation using the spectral envelope described in Eq.(3.24) as the filter frequency response. As the spectral envelope described in Eq.(3.24) is not a rational function it can not be implemented directly. An approximation for it is given by the Mel log spectrum approximation (MLSA) filter proposed in Fukada et al. (1992) for $\gamma = 0$ and the Mel-generalized log spectrum approximation (MGLSA) filter in Tokuda et al. (1994).

The excitation signal can also be obtained simply by the generated F_0: that is, the so called simple excitation, instead of a mixture. For voiced segments, the excitation signal is a pulse and for unvoiced segments, white noise.

3.2 Acoustic model

This section will describe how to train HMMs using the extracted acoustic parameters described in the previous section. We also present how the sequence of spectral envelope and excitation parameters that feed the synthesis mechanism of the vocoder are generated from HMM-based acoustic models. Additionally, we show how to adapt acoustic models to other speakers and speaking styles and methods proposed to alleviate oversmoothing of vocoded parameters due to statistical modelling.

3.2.1 Hidden Markov models

HMMs are generative models that represent data through a sequence of states of a Markov chain. The Markov chain defines a discrete and finite state space, the number of states is a design decision. HMMs are widely used to represent time series of data where modelling data in sequential states seems like a natural choice. The Markov property implies a limitation to the model as there is an assumption that the future state depends only on the current one. In each state, data is represented by a distribution that can be, for instance, a Gaussian mixture model (GMM) or a more complex distribution such as a deep belief network or a restricted Boltzmann machine. HMMs can be classified as discrete or continuous, depending on the type of data they are representing. In the first case, discrete – categorical – data is represented by discrete probabilities associated with each state while in the second, the state output probabilities are continuous distributions in the form of probability density functions. The "hidden" term in HMM refers to the fact that the state sequence is an unknown variable, which in practice means that the parameters of an HMM are estimated by marginalizing the objective function across all possible state sequences.

The three core problems associated with HMMs are:

– Efficient evaluation of the marginal over all states: how to compute the probability density of an observation sequence given an HMM, the so-called likelihood – how likely the data is to be generated by that model. Method: the forward-backward algorithm.

– Model parameter estimation: how to estimate the parameters that define an HMM given a training dataset. Method: the expectation maximization algorithm.

– Computation of the optimal state sequence: given the observation sequence how to find the most probable state sequence. Method: Viterbi algorithm.

A comprehensive tutorial on how these problems are tackled is given in Rabiner (1989). In the next section we present how to train HMMs following the derivations described in Rabiner (1989).

An HMM is defined by the number of states it contains, the state transition probability distribution matrix $\boldsymbol{A} = \{a_{ij}\}$ which holds the probability of transitioning from state i to state j, the emission probability distribution for each state, referred to as the output probability distribution $\boldsymbol{B} = \{b_i(.)\}$ and finally the probability of a state being

the initial state of a sequence, referred to as the initial state probability $\mathbf{\Pi} = \{\pi_i\}$. It is common to see an HMM being represented with the following notation:

$$\lambda = (\boldsymbol{A}, \boldsymbol{B}, \mathbf{\Pi}) \tag{3.30}$$

When the state output probability distribution $b_i(.)$ is modelled by a single multivariate – the output is a vector rather than a scalar – Gaussian distribution, we can write the following:

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \tag{3.31}$$

$$= \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}_i|}} \exp\left\{ -\frac{1}{2} (\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\boldsymbol{o}_t - \boldsymbol{\mu}_i) \right\} \tag{3.32}$$

where $\boldsymbol{o}_t$ is the observation vector of dimension d consisting of parameters extracted by the vocoder at frame t and $\top$ is the transpose operator. The parameters that define the Gaussian distribution are: the Gaussian mean vector $\boldsymbol{\mu}_i$ of dimension dx1 and its covariance matrix $\mathbf{\Sigma}_i$ of dimension dxd.

In the next sections we will show how to train HMMs, how to train HMMs as synthesis models and how to generate parameters from them, assuming always that the state output probability distribution is modelled by one Gaussian only.

3.2.2 Training HMMs

Training an HMM is the process of finding the parameters that define it, given a training dataset. One way of finding the model parameters is to maximize the likelihood of the training data acoustic parameters $\boldsymbol{O}$ given the model λ:

$$\lambda_{\max} = \arg\max_{\lambda} p(\boldsymbol{O} \mid \lambda) \tag{3.33}$$

where $\boldsymbol{O} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..\boldsymbol{o}_T^\top,]^\top$ and T is the parameter sequence length.

The maximization of the likelihood when the state sequence is unknown is obtained by applying the expectation maximization (EM) algorithm, also called the Baum-Welch algorithm when applied to HMMs. The EM algorithm is a method for maximizing the likelihood of distributions that depend on hidden variables, i.e. whose values are not available in a supervised fashion. The basic idea behind the EM algorithm is that the hidden variable, in this case the state sequence, can be eliminated by marginalization. When that is done, the maximization of the log-likelihood is achieved by the maximization of the so-called auxiliary function, defined as:

$$Q(\lambda^\kappa, \lambda^{\kappa+1}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{O}, \lambda^\kappa) \log p(\boldsymbol{O}, \boldsymbol{q}|\lambda^{\kappa+1}) \tag{3.34}$$

where $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence and the index κ refers to the iteration, so λ^{κ} and $\lambda^{\kappa+1}$ are the models of the current and the subsequent iteration.

In the EM algorithm, the maximization of the auxiliary function is obtained iteratively. Given a model initialization, in each iteration the auxiliary function is estimated and then maximized providing the model for the next iteration. The maximization is achieved through the calculation of the probabilities of state occupancy and transition given the observation vector sequence and the model posterior probabilities. The posterior probabilities $\gamma_t(i)$, the probability of being in state i at time t and $\xi_t(i,j)$, the probability of transitioning from state i to state j at time t, are given by:

$$\gamma_t(i) = p(q_t = i|\boldsymbol{O}, \lambda) \tag{3.35}$$

$$= \frac{\alpha_t(i)\,\beta_t(i)}{p(\boldsymbol{O}|\lambda)} \tag{3.36}$$

$$= \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\,\beta_t(j)} \tag{3.37}$$

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j|\boldsymbol{O}, \lambda) \tag{3.38}$$

$$= \frac{\alpha_t(i)\,a_{ij}\,b_j(\boldsymbol{o}_{t+1})\,\beta_{t+1}(j)}{\sum_{j=1}^{N} \alpha_t(j)\,\beta_t(j)} \tag{3.39}$$

where:

$$\alpha_t(i) = p([\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_t]\ ,\ q_t = i|\lambda) \tag{3.40}$$

$$\beta_t(i) = p([\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, ..., \boldsymbol{o}_T]\ ,\ q_t = i|\lambda) \tag{3.41}$$

$\alpha_t(i)$ is the so called forward and $\beta_t(i)$ the backward probability.

The forward and backward probabilities can be calculated by following the given relation:

$$p(\boldsymbol{O}, \boldsymbol{q}|\lambda) = p(\boldsymbol{O}|\boldsymbol{q}, \lambda)P(\boldsymbol{q}|\lambda) \tag{3.42}$$

$$= \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{o}_t) \tag{3.43}$$

The calculation of the posteriors from the estimated model parameters allows for the maximization step of EM that will then update the model parameters. Considering the state output probability distribution made of a single Gaussian, we have the

following update equations:

$$\pi_i = \gamma_1(i) \tag{3.44}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.45}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.46}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{t=1}^{T} \gamma_t(i) (\boldsymbol{o}_t - \boldsymbol{\mu}_i)(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^\top}{\sum_{t=1}^{T} \gamma_t(i)} \tag{3.47}$$

3.2.3 Training synthesis models

To obtain models that can generate acoustic sequences to drive a vocoder and synthesize high quality speech, the standard HMM training as used for automatic speech recognition has been reformulated to account for the following requirements: additional acoustic features, explicit duration modelling and a full-context dependency. We will see in the next few sections how each of these requirements have been met for the training of synthesis models.

3.2.3.1 Feature vectors and state emission probabilities

A typical observation vector is constructed from the vocoded parameters by a concatenation of their static values plus their dynamics, represented by delta $\Delta \boldsymbol{c}_t^\top$ and delta-delta $\Delta^2 \boldsymbol{c}_t^\top$ values as follows:

$$\boldsymbol{o}_t = [\boldsymbol{c}_t^\top, \Delta \boldsymbol{c}_t^\top, \Delta^2 \boldsymbol{c}_t^\top]^\top \tag{3.48}$$

where $\boldsymbol{c}_t$ is a static parameter of order M with the vocoded parameters extracted for analysis window t and $\boldsymbol{o}_t$ the observation vector of dimension $3M$x1. The dynamics are calculated from the statics as follows:

$$\Delta^{(n)} \boldsymbol{c}_t = \sum_{\tau=-L^{(n)}}^{L^{(n)}} w_\tau^{(n)} \boldsymbol{c}_t \qquad 0 \leqslant n \leqslant 2 \tag{3.49}$$

where $2L^{(n)} + 1$ is the size of the window used to calculate the dynamics of order n, $L^{(0)} = 0$ and $w_0^{(0)} = 1$.

As presented previously, the synthesis module of the vocoder requires parameters that describe both the spectral envelope and the excitation signal. To drive the generation of the excitation signal, F_0 values need to be modelled. Unlike the spectral

and aperiodicity parameters, F_0 is not strictly continuous. For voiced segments, F_0 is continuously defined but for unvoiced segments it is undefined, however, this does not mean that it takes the value of zero. One of the ways of handling this is to consider F_0 as a multispace variable (Tokuda et al., 2002), where one space assumes continuous values and the other space a discrete distribution. For each state, there is a label that indicates which space F_0 is and which distribution is attributed to it.

In order to maintain synchronization between the different parameters (spectral, aperidiodicity and F_0), the observation vector adopted in TTS is composed of multiple separate streams in a multistream HMM (Young et al., 2006). Each stream contains static and dynamic representations of this data. Each stream refers here to sections of the observation vector that are considered to be statistically independent of each other. When the observation vector is made up of more than one stream, the output probability is the product of the probability of each stream as follows:

$$b_i(\boldsymbol{o}) = \prod_{s=1}^{S} b_{is}(\boldsymbol{o}_s) = \prod_{s=1}^{S} \left\{ w_{is} \mathcal{N}(\boldsymbol{o}_s; \boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}_{is}) \right\} \tag{3.50}$$

where S is the number of streams, w_{is} is a weight associated with the output probability of state i and stream s and $\mathcal{N}(\boldsymbol{o}_s; \boldsymbol{\mu}_{is}$ the output probability associated with stream i and state s modelled as a single Gaussian. Multistream training keeps the synchronization of spectral and excitation models while still allowing them to be separately tied, as we will soon discuss.

3.2.3.2 Duration modelling

Without explicit duration modelling, the state duration of an HMM would be given by the distribution of the transition probabilities which in turn give an exponential decaying distribution. As this is not a good model to generate natural sounding phone durations, explicit duration modelling in the form of the semi-Markov structure (Ferguson, 1980; Zen et al., 2007c) was proposed. Under this framework, the duration of state k is modelled by a Gaussian distribution:

$$p_k(d_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(d_k - \mu_k)^2}{2\sigma_k^2} \right) \tag{3.51}$$

where d_k, μ_k and σ_k are the duration of state k, the mean and the variance of the duration distribution of state k.

3.2.3.3 Context dependency and parameter tying

In automatic speech recognition, a triphone defines an HMM model: a different model is trained for a phone segment depending on the phonemes that come before and after. Although using triphone context might be sufficient for the decoding of text from acoustics, to characterize high quality natural sounding speech, a wider context is required. Context dependency for speech synthesis should take into account, in addition to phoneme assignment, information about lexical stress, pitch accent, tone and part-of-speech, i.e. linguistic information that can potentially influence the acoustic realization. A richer context can potentially create a vast amount of models which means in practice some contexts will not have any or very few example in the training set. Consequently, it is possible that at generation time unseen contexts will appear. The number of possible contexts have to be restricted. It is however not clear which linguistic specification is sufficient to define an acoustic model. Rather than creating rules for what is a relevant context, such as the triphone rule used for speech recognition, in statistical parametric synthesis the underlying context dependency is found automatically (Yoshimura et al., 1999).

HMM training is carried out first for each monophone in context-independent training, creating context-independent HMMs (CI-HMMs). The CI-HMMs are then tied together using a stream-dependent tree-based state clustering: a different decision tree will be built for spectral, excitation and duration parameters. The streams of spectral, excitation and duration parameters are clustered independently with the assumption that their dependency on the linguistic context will be different: F_0 and duration are more affected by supra-segmental linguistic specification while spectral parameters are affected by localised linguistic characteristics like the phone. Each leaf of the decision tree refers to a context-dependent state (CD-HMMs). The linguistic specification determined by the questions leading to the leaf nodes indexes the CD-HMMs. The questions associated with the decision trees in practice define regions in the linguistic space (the multidimensional space covered by all possible linguistic specifications) so an unseen specification will be associated with the model of the region it comes from. In other words, any context will reach one of the leaf nodes, from the root node then selecting the next node depending on the answer about the current context. In the clustering technique, the size of the decision tree is automatically controlled based on the minimum description length criterion (Shinoda and Watanabe, 2000).

3.2.4 Parameter generation

In the synthesis part, the text to be synthesized is converted by the front-end to a sequence of linguistic specifications. Each linguistic specification will drive the selection of a CD-HMM by answering the linguistic questions of the decision trees. Given this concatenated sequence of CD-HMMs, the utterance HMM λ is constructed. The most likely observation sequence is given by the maximization of the likelihood function $P(\boldsymbol{O}|\lambda)$. There is no known method to find a closed form solution to this maximization problem. Tokuda et al. (2000) show how to find the solution iteratively by using the EM algorithm. Tokuda et al. (2000) also show that a closed form solution can be found if we consider just the most likely state sequence:

$$\begin{aligned}
\boldsymbol{O}_{max} &= \arg\max_{\boldsymbol{O}} p(\boldsymbol{O}|\lambda) && (3.52)\\
&= \arg\max_{\boldsymbol{O}} \sum_{\text{all } \boldsymbol{q}} p(\boldsymbol{O},\boldsymbol{q}|\lambda) && (3.53)\\
&\backsimeq \arg\max_{\boldsymbol{O}} \max_{\boldsymbol{q}} p(\boldsymbol{O},\boldsymbol{q}|\lambda) && (3.54)\\
&= \arg\max_{\boldsymbol{O}} \max_{\boldsymbol{q}} p(\boldsymbol{O}|\boldsymbol{q},\lambda)P(\boldsymbol{q}|\lambda) && (3.55)
\end{aligned}$$

where, following the definitions from Section 3.2.2, $\boldsymbol{q} = \{q_1, q_2, \ldots, q_T\}$ is a state sequence and $\boldsymbol{O} = [\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, ..\boldsymbol{o}_T^\top,]^\top$ is the speech parameter observation sequence column vector with length $3MT$, M being the length of the static vector $\boldsymbol{c}_t$ and T the number of states.

The problem can be divided into two maximizations:

$$\begin{aligned}
\boldsymbol{q}_{max} &= \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) && (3.56)\\
\boldsymbol{O}_{max} &= \arg\max_{\boldsymbol{O}} p(\boldsymbol{O}|\boldsymbol{q}_{max},\lambda) && (3.57)
\end{aligned}$$

Assuming the HMM state transition goes from left-to-right without skipping states, it is possible to find the state sequence from the model by using the state duration probability, which in the HSMM paradigm is explicitly modelled by a distribution:

$$\begin{aligned}
\boldsymbol{q}_{max} &= \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda) && (3.58)\\
&= \arg\max_{\boldsymbol{q}} \prod_{k=1}^{K} p_k(d_k) && (3.59)
\end{aligned}$$

where $p_k(d_k)$ is the probability of duration d_k in state k, i.e. the probability that a segment of d_k duration is emitted from state k; K is the number of states visited during the duration of T – given by the model specification. The total duration has to

be achieved so $\sum_{k=1}^{K} d_k = T$. For the Gaussian distribution, the state durations that maximize Eq.(3.59) is given by (Yoshimura et al., 1998):

$$d_k = \mu_k + \rho\, \sigma_k^2 \tag{3.60}$$

$$\rho = \left(T - \sum_{k=1}^{K} \mu_k \right) \Big/ \sum_{k=1}^{K} \sigma_k^2 \tag{3.61}$$

where μ_k and σ_k are the mean and the variance of the duration distribution of state k and ρ is a parameter that can be controlled by a desired total duration as we will soon show. From the values of d_k, we know how many frames are emitted by each state and therefore the state sequence. When synthesizing a sentence it is possible to set a desired total duration T. From the equations above, we are able to calculate the state duration d_k, however, rounding errors in the process of approximating a real value (time) to a integer value (number of states) means that the generated sentence will not necessarily have exactly duration T.

It is possible to control the speaking rate by changing the value of ρ. When ρ is set to zero then $\sum_{k=1}^{K} \mu_k = T$, if ρ is negative then total duration is smaller – faster – and if positive overall duration if longer – slower rate. Alternatively, to control the speaking rate by changing ρ we can define a scaling factor φ:

$$\rho = \left(\varphi \sum_{k=1}^{K} \mu_k - \sum_{k=1}^{K} \mu_k \right) \Big/ \sum_{k=1}^{K} \sigma_k^2 \tag{3.62}$$

where $\varphi = 1$ sets ρ to zero and the normal speaking rate, $\varphi > 1$ slows down the rate and $\varphi < 1$ speeds it up.

Given that the state sequence is now known, we can proceed to find the observation sequence $\boldsymbol{O}_{max}$:

$$\boldsymbol{O}_{max} = \arg\max_{\boldsymbol{O}} p(\boldsymbol{O}|\boldsymbol{q}_{max}, \lambda) \tag{3.63}$$

$$= \arg\max_{\boldsymbol{O}} \mathcal{N}(\boldsymbol{O}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \tag{3.64}$$

Without any constraints, the maximization in Eq.(3.64) will set $\boldsymbol{O}_{max}$ to be the sequence of the Gaussian mean vectors. To create parameters whose temporal trajectories are not a sequence of abrupt transitions through the mean vectors of the sequence of HMMs, a constraint needs to be added. This constraint is given by the relation between the static and dynamic components that define the observation vector, seen in Eq.(3.48). The following holds:

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C} \tag{3.65}$$

where:

$$\boldsymbol{C} = [\boldsymbol{c}_1^\top, \boldsymbol{c}_2^\top, ..\boldsymbol{c}_T^\top,]^\top \tag{3.66}$$

$$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_T]^\top \otimes \boldsymbol{I}_{MxM} \tag{3.67}$$

where $\otimes$ is the tensor product, $\boldsymbol{I}_{MxM}$ the identity matrix of dimension MxM, $\boldsymbol{W}$ is a $3T$xT matrix and $\boldsymbol{W}_t$ a $3T$x1 vector constructed from the weights w seen in Eq.(3.49) which define how the delta values are calculated from the static values:

$$\boldsymbol{W}_t = [\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}] \tag{3.68}$$

$$\boldsymbol{w}_t^{(0)} = [\boldsymbol{0}_{t-1}, 1, \boldsymbol{0}_{T-t}]^\top \tag{3.69}$$

$$\boldsymbol{w}_t^{(1)} = [\boldsymbol{0}_{t-L^{(1)}-1}, w_{-L^{(1)}}^{(1)}, ..., w_0^{(1)}, ..., w_{L^{(1)}}^{(1)}, \boldsymbol{0}_{T-L^1-t}]^\top \tag{3.70}$$

$$\boldsymbol{w}_t^{(2)} = [\boldsymbol{0}_{t-L^{(2)}-1}, w_{-L^{(2)}}^{(2)}, ..., w_0^{(2)}, ..., w_{L^{(2)}}^{(2)}, \boldsymbol{0}_{T-L^2-t}]^\top \tag{3.71}$$

where $\boldsymbol{0}_k$ is a kx1 vector of zeros.

With the constraint of Eq.(3.65), the maximization in Eq.(3.64) becomes:

$$\begin{aligned} \boldsymbol{C}_{max} &= \arg\max_{\boldsymbol{C}} \mathcal{N}(\boldsymbol{WC}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) && (3.72) \\ &= (\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} && (3.73) \end{aligned}$$

$$\boldsymbol{\Sigma} = diag[\boldsymbol{\Sigma}_{q_1}^{-1}, \boldsymbol{\Sigma}_{q_2}^{-1}, ..., \boldsymbol{\Sigma}_{q_T}^{-1}]^\top \tag{3.74}$$

where:

$$\boldsymbol{\mu} = diag[\boldsymbol{\mu}_{q_1}^{-1}, \boldsymbol{\mu}_{q_2}^{-1}, ..., \boldsymbol{\mu}_{q_T}^{-1}]^\top \tag{3.75}$$

Rewriting Eq.(3.73), the generated parameter sequence is given by:

$$\boldsymbol{C}_{max} = \boldsymbol{R}^{-1} \boldsymbol{r} \tag{3.76}$$

where:

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W} && (3.77) \\ \boldsymbol{r} &= \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} && (3.78) \end{aligned}$$

The parameter generation algorithm has a closed form as in Eq.(3.76) in the case where the state sequence is assumed independent of the observation sequence, and thus can be found (or provided externally) in advance. As previously stated, a more complex iterative solution using EM is required if this independence assumption is not made and it is desired to jointly optimise for state sequence and observation sequence.

By utilizing the constraints on the dynamics for parameter generation, the trajectories of the parameters over time are smoother (rather than being a sequence of means) which in turn increases the quality of the generated speech. As we saw in the previous section, this constraint was not used at training time: the HMMs are trained as if the static and dynamic components of the observation vector are independent of each other. To include the explicit relation $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{c}$ and correct for the training/generation inconsistency Zen et al. (2007b) proposed a reformulation of the training algorithm referred to as the trajectory HMM. Trajectory HMMs can accommodate the dynamic constraints with no additional model parameters.

3.2.5 Adaptation

The statistical framework enables the use of model adaptation techniques that can adjust trained models in such a way that they better describe a new dataset. Adaptation methods for HMMs in speech technology were first proposed in the context of speech recognition in order to adapt a system for a particular speaker, channel condition or language (Gauvain and Lee, 1994; Digalakis et al., 1995; Leggetter and Woodland, 1995). For TTS, adaptation techniques are used, for instance, to create voices for a particular speaker from a small amount of speech data through the use of a model trained with a large amount of data from other speakers – the average voice model (Yamagishi, 2006). Speaker adaptation can also be used to create voices with different speaking styles, which could be emotional states like happy and sad or production-related styles like hyper and hypo articulation, clear speech and noise-driven Lombard speech. In this context, it is possible to adapt a model trained with neutral speech data to one of these styles using a small amount of style-matched training data (Yamagishi et al., 2004; Picart et al., 2011; Raitio et al., 2011a)

The main techniques for adaptation of HMMs are based on linear regression applied to the model parameters: Gaussian means and covariance matrices are adjusted using a linear transform. Maximum likelihood linear regression (MLLR) adaptation updates the mean vectors in order to maximize the likelihood of the data used for adaptation (Leggetter and Woodland, 1995; Tamura et al., 2001). The same transform is shared across different states because the limited adaptation data might not cover all models. MLLR performs linear transforms of mean vectors of the state output probability distributions in the following way:

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t \; ; \; \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_k \; , \; \boldsymbol{\Sigma}_i) \tag{3.79}$$

where $\boldsymbol{\zeta}_k$ and $\boldsymbol{\varepsilon}_k$ are a dxd matrix and a d-dimensional vector, respectively. k denotes k-th regression class.

In addition to adapting the means, one can also adapt the covariance in the models using the same matrices, this is called the constrained MLLR (CMLLR) (Gales, 1998). The state output probability distribution is affine-transformed as follows:

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t \; ; \; \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_k \; , \; \boldsymbol{\zeta}_k \boldsymbol{\Sigma}_i \boldsymbol{\zeta}_k^\top) \tag{3.80}$$

Linearly transforming the mean and covariance parameters of the Gaussian distribution can be translated to a linear transformation of the observation vector in the following way (Gales, 1998):

$$b_i(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t \; ; \; \boldsymbol{\zeta}_k \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_k \; , \; \boldsymbol{\zeta}_k \boldsymbol{\Sigma}_i \boldsymbol{\zeta}_k^\top) \tag{3.81}$$

$$= |\boldsymbol{\zeta}'| \; \mathcal{N}(\boldsymbol{\zeta}'_k \boldsymbol{o}_t - \boldsymbol{\varepsilon}'_k \; ; \; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.82}$$

where $\boldsymbol{\zeta}' = \boldsymbol{\zeta}^{-1}$ and $\boldsymbol{\varepsilon}'_k = \boldsymbol{\zeta}^{-1} \boldsymbol{\varepsilon}_k$. A linear transformation applied to the Mel cepstral coefficients can be transposed to a linear transformation of the model parameters.

The transforms $[\boldsymbol{\zeta}; \boldsymbol{\varepsilon}]$ can be estimated using different criteria like MLLR, CMLLR and maximum a posterior (MAP)-based criteria like the structured MAP linear regression (SMAPLR) and the constrained structural MAP linear regression (CSMAPLR) (Yamagishi et al., 2009). Given enough data for adaptation, it is useful to combine linear regression adaptation with MAP adaptation (CSMAPLR-MAP) (Yamagishi et al., 2009).

Additionally, it is possible to combine LR-based model adaptation with other types of model transformation such as vocal tract length normalization (VTLN) (Saheer et al., 2012). The transformation applied in VTLN is done by one parameter only, the warping factor. As it is a much simpler transformation, VTLN requires very little adaptation data but does not capture as much of the speaker characteristics as linear transformation based adaptations do. Using VTLN is especially attractive for TTS because transformation can be done with no additional parameters when the MGC spectral parametrization, described in Section 3.1.3.2, is adopted (Pitz and Ney, 2005; Saheer et al., 2010). The MGC parametrization enables spectral envelope warping via the parameter α. This means that the relation between MGCs of a linear and a warped spectral envelope can be described by α, so no additional model parameters needs to be added.

3.2.6 Oversmoothing

The parameters generated by the acoustic model to synthesize speech are the mean vectors of the Gaussian models, which in turn are calculated by averaging observation vectors extracted at training time as seen in Eq.(3.46). This averaging creates an over-smoothed spectral envelope, the rich details of peaks and valleys is lost. To overcome this problem, a few methods have been proposed: post filtering (Ling et al., 2006), Global Variance (GV) (Toda and Tokuda, 2005, 2007) and minimum generation error training (MGE) (Wu and Wang, 2006).

A simple solution to create less smoothed spectral envelopes is to modify the generated spectral parameters. One possible way of enhancing spectral sharpness is to bring line spectral pairs closer to each other as proposed by Ling et al. (2006). We will see a fuller description of this in Chapter 4. Although it can increase the quality of generated speech, it is done independently of the naturally occurring spectral peaks, so such post filtering can introduce audible artefacts.

A different way of looking at the problem is to model the variance observed in the training stage and reproduce it at generation time. The global variance (GV) method (Toda and Tokuda, 2005, 2007) is used to counteract the over smoothing problem by modelling the GV – variance calculated across a utterance – of the parameters extracted during training. At training time, the GV vectors – the variance calculated for each dimension of the observation vector – are calculated for each utterance in the training corpus and a single multivariate Gaussian distribution is used to model them. The maximum likelihood training then considers the maximization of the likelihood of both the model trained with vocoded parameters and the GV model. At generation time, the observation vector is obtained by maximizing the likelihood of the HMM model *and* the GV model. The additional term makes it hard to find a closed form solution, so Toda and Tokuda (2007) proposed an iterative method based on gradient descent. The solutions for the first order and second order gradient descent are presented in Toda and Tokuda (2007).

GV remediates oversmoothing by modelling variance explicitly. Alternatively, one can improve one of the causes of oversmoothing: the training criterion. Acoustic model training using ML does not consider the mismatch between generated and natural parameters. To fix this problem the Minimum Generation Error (MGE) HMM training method was proposed (Wu and Wang, 2006). MGE training aims at minimising the generation error of the modelled acoustic parameters during training. In that way,

the training procedures incorporate the parameter generation for the error calculation. An appropriate choice of error criterion guides the quality of the generated acoustic features. In its simpler form, MGE training operates on the Viterbi-aligned HMM sequence, initialised by the results of ML training. The first proposed optimization function (Wu and Wang, 2006) for MGE training was simple Euclidian distance between original and generated line spectral pairs (LSP). Since then, other versions of MGE have been proposed: log spectral distance from LSPs (Wu and Tokuda, 2008) and from the STRAIGHT extracted spectrum (Wu and Tokuda, 2009).

3.3 Subjective evaluation

It is common to illustrate the performance of HMM-based speech synthesis by presenting spectral distortion measures like cepstrum distance and model distortion measures like the likelihood of the training set (Zen et al., 2007b; Yamagishi et al., 2009; Ling and Dai, 2012). Although they serve as an indication of how well the synthesis model represents natural speech, measuring the quality of synthetic speech automatically is a challenge even when a reference natural speech signal is available, which is most likely not the case (Möller and Falk, 2009). Although non-intrusive measures, measures that do not require a reference speech signal, have been proposed for TTS (Falk et al., 2008; Norrenbrock et al., 2012) for a true measure of quality, listening tests remain the gold standard.

3.3.1 Procedure

Through listening tests, it is possible to assess different quality dimensions by asking listeners to perform different tasks. Listening tests can provide scores for naturalness, speaker similarity, expressiveness and intelligibility in both quiet and adverse conditions. When designing a listening test it is important to consider all factors that may affect the scores. It is thus very important to select stimuli and presentation method carefully and to test a sufficient number of listeners. For instance, a poor quality synthetic voice will likely be judged worse if better voices are present in the same test even when participants are not explicitly asked to compare them. The same effect can occur when including both natural speech and synthetic speech in the same test. As each listener and test generates different scores, comparing between tests is potentially "dangerous", unless due care is taken. Listener variability is also a factor that needs

to be controlled, for example having native and non native speakers in the same test can create conflicting results as it was noted that these two listeners types perceive speech in clean (Flege et al., 1997) and in noise (van Wijngaarden et al., 2002) very differently. It is better to first treat their data separately and point out the differences between them before averaging scores. Another factor that needs to be accounted for is listening impairment. A basic hearing test can be done that assesses whether participants can detect a pulse signal emitted at a certain level (30 or 25dB depending on the criteria) at a range of frequencies bilaterally. One fails the test if two different frequencies are not detected by the same ear or if the same frequency is not detected by both ears. Participants that fail the screening are still asked to take the listening test as they can not be informed, for ethical reasons, whether they passed the test or not. Their scores are however not counted.

3.3.2 Types of listening test

The most common type of task is to ask participants to judge how natural a stimulus is on a scale of one to five, the average of these ratings is then taken to generate the mean opinion score (MOS). In the field of speech code evaluation this kind of test is also referred to as the absolute category rating (ACR) test as quality judgements are made without any reference. To compare speech signals with small perceptual differences, the most appropriate method is a direct comparison test. In degradation category rating (DCR) tests participants listen to a reference signal (non degraded) as well and are asked to rate (on a scale from one to five) the perceived degradation in quality with regards to this signal. To allow for a random ordering of presentation, comparison category rating (CCR) tests ask listeners to rate the second stimulus compared to the first one on a scale that ranges from minus 3 (much worse) to 3 (much better). Additionally, the so-called MUSHRA (multi-stimulus test with hidden reference and anchors) test allows for the comparison of multiple stimuli. Participants are presented with several speech samples and a known reference which should not be rated. Each stimuli is then rated on a continuous scale defined from 0 (lowest quality) to 100 (i.e. best quality). The scores quantify the quality degradation of each stimuli under test compared to the reference signal. It is also possible to judge naturalness by preference tests using the AB test: is stimuli A more natural than stimulus B, and the ABX test: is stimuli X perceptually closer to A or to B. Comparison tests and MOS can also be used to measure speaker similarity. The MOS similarity task can also be viewed as a comparison

test as participants are asked to scale similarity to a particular speaker. These are the most common type of tests, and although they can tell us about the general quality of a voice, they do not provide segment-level judgements.

For the purposes of this thesis, it is intelligibility that we wish to assess. There are a number of tests available for testing intelligibility, a few examples of which we will explain here. Lexical decision tasks involve identifying whether the sample heard is a dictionary word (eg. coloured) or a nonword (eg. coobered). Word recognition tasks require the participant to identify which word was spoken or at which point a word becomes identifiable - known as the gating paradigm. Word recall tasks involve asking participants which word or sentence they heard to test for instance the effect on memory and cognitive overload. We chose to use a transcription test, which requires the participant to type in what they hear. The results can then be compared to the text used to generate the stimuli to provide a measure of the intelligibility through word accuracy rates (WAR) or phoneme error rates. A design decision crucial to transcription tasks is the choice of sentences or words. Examples used in our work include Matrix sentences (Dreschler, 2006), semantically unpredictable sentences (SUS) (Benoit, 1990) and Harvard sentences (IEEE, 1969). Other sentence types exist and some tests require the design of specific sentence or word material. A common sub-version of the transcription test is the modified rhyme test (MRT) (Fairbanks, 1958), which although not used explicitly in this thesis, has been used extensively in the research that helped inform it. MRT uses a set of 300 words organised in groups of six, where words in a group differ from each other at only one position. Participants listen to a subset of the words and must either report which word they heard (open set) or choose the word from the list of six possibilities (closed set).

3.3.3 Intelligibility studies

Extensive work has been done on measuring the intelligibility of synthetic speech and a good survey can be found in Winters and Pisoni (2006). Knowing in which circumstances synthetic speech is less intelligible can help in the design of better TTS systems as well as providing insights into how speech is perceived. In particular, speech perception researchers want to know what the advantages of natural speech are over synthetic speech. The main body of work done in speech perception of TTS has been performed using formant-based TTS voices. In comparison, very few studies have evaluated the intelligibility of concatenative and HMM-based systems. Studies with

both formant and concatenative systems all point to the same conclusion: compared to natural speech, synthetic speech is harder to perceive and therefore less intelligible (Winters and Pisoni, 2006). This result has been observed in many different studies that vary by the choice of TTS system, sentence material, listening condition, task and listener group.

3.3.3.1 Segment-level

Early studies, e.g. (Nye and Gaitenby, 1973; Pisoni et al., 1985; Logan et al., 1989), investigated segmental intelligibility errors using MRT. To provide greater detail, the majority of these studies report not only word error rates but also phoneme error rates, organized in groups of phonemes. Results of these studies from both open and closed set MRTs showed higher error rates for rule-based synthetic speech than for natural speech (Logan et al., 1989). One issue with MRTs for intelligibility testing is their relative simplicity. Small numbers of potentially confusable words mean that with natural speech and good quality synthesizers a ceiling is inevitably reached, albeit less quickly in open set conditions.

Testing intelligibility in noise reduces the likelihood of reaching this ceiling. Intelligibility in noise is of course of interest in its own right as it reflects more realistic listening conditions. In noise, both synthetic and natural speech become less intelligible, however the intelligibility degradation when noise levels increase is significantly worse for TTS voices (Koul and Allen, 1993). MRT evaluations reveal not only more instances of errors found in clean conditions but also new phoneme confusions for synthetic speech (Winters and Pisoni, 2003). As might reasonably be expected, the potential confusions are dependent on the type of speech and the type of noise.

More recently, a study compared formant and concatenative TTS systems (Venkatagiri, 2003) by evaluating segmental intelligibility in noise – multitalker babble – using the MRT words in a open set fashion. This study found that all TTS methods tested were significantly less intelligible than natural speech at both SNRs tested, by at least 22%. Additionally, the study found that concatenative systems created more vowel confusions while the formant-based system gave more consonant confusions, which indicates that concatenative techniques are better at modelling consonants while formants are better at modelling vowels. MRT can lead to useful findings, but alone it is not sufficient because it provides quite a limited number of confusions and the phoneme distribution across the test is not balanced. More complex methods are also required because MRT is valid only at the single word or phoneme level.

3.3.3.2 Sentence level

For greater understanding, it is necessary to move beyond isolated words to look at complete sentences. Results of sentence tasks can help us identify the semantic and sentential factors in the perception of speech. It has been shown that perception of synthetic words in a sentence is significantly better than in isolation (Hoover et al., 1987; Mirenda and Beukelman, 1987). For poor quality synthesizers it was found that the semantic context gives more improvement than moving from isolation to a sentence – the sentential context (Hoover et al., 1987). For high quality TTS voices, an improvement caused by sentential context was also observed (Mirenda and Beukelman, 1987).

Due to its relative newness there have been far fewer segmental studies performed for the perception of HMM-based systems, but a substantial body of work that has been done with sentences on the evaluation of corpus-based TTS systems originated with the introduction of the annual Blizzard challenge in 2005 (Black and Tokuda, 2005). HMM-based TTS entries featured and have continued to do so in every subsequent challenge including a baseline provided by the challenge organisers. The challenge evaluates different TTS systems trained with the same database. Scores are obtained using the same text material, listening conditions and listeners, to provide a fair comparison of the different TTS systems. Listening tasks judge naturalness in MOS scores, similarity and intelligibility in quiet and noise. The first few challenges 2005-2009 evaluated intelligibility of sentences in quiet, some without the natural speech baseline for comparison. Some systems obtained intelligibility scores comparable to natural speech, including HMM-based entries e.g. (Yamagishi et al., 2008b). Overall the intelligibility differences between the synthetic voices and natural speech ranged between 5% to 40% across the entries. In 2010, the Blizzard challenge was expanded to include the evaluation of intelligibility in noise. Overall intelligibility was seen to drop faster for synthetic speech (King and Karaiskos, 2010). Synthetic speech can however be more intelligible than natural speech in noisy situations when the synthesizer is modified as in the Blizzard 2010 entry described by Suni et al. (2010). In this entry the formant structure was enhanced via a post-filtering technique during the open phase of the glottal cycle, pitch was raised and the spectral tilt was halved via modifications to the glottal pulse harmonic structure, a combination that resulted in large intelligibility gains in noise.

It is also of interest to evaluate TTS at higher speaking rates. Intelligibility at higher speaking rates is important to people with visual disabilities – a large proportion

of TTS users – as they tend to listen to TTS at faster speeds because they are expert users. A large scale test has recently evaluated the intelligibility of eight different TTS voices across a range of speaking rates (Syrdal et al., 2012). Two systems of each type: formant, diphone, unit selection and HMM-based synthesis were tested using SUS at 200-450 words per minute in quiet conditions. Significant differences were found between the systems, with unit selection systems achieving higher scores across all rates. Errors increase significantly with speech rate. Contrary to what was found in studies with visually impaired people (Stent et al., 2011), this study found the unit selection systems to be more intelligible than formant synthesizers at higher rates. The HMM system was as intelligible as the unit selection system at the human default speaking rate of 200 words per minute.

3.3.3.3 Special cases

Organizing traditional listening experiments can be costly, complex and time consuming, especially if a large number of participants is required or the target language is not the native language spoken where the test is taking place. Wolters et al. (2010) investigated whether Amazon mechanical turk (AMT) can be used to compare the intelligibility of speech synthesis systems. AMT is a platform provided by Amazon that allows tasks to be crowdsourced. Participants are recruited and paid through Amazon and do the task using their own computer. It is impossible to control the listening conditions but there are some ways of checking whether listeners are performing the task appropriately. The word error rates for SUS in quiet conditions found by Wolters et al. (2010) using AMT are much worse than those obtained in controlled lab conditions but the relative differences between the systems are similar. Findings such as the ones reported in Wolters et al. (2010) have led to the adoption of AMT by many research groups.

Whereas most previous evaluations have required participants to use headphones and often soundproof booths, Raitio et al. (2012) evaluated the intelligibility of HMM-based TTS in noise in a more realistic listening environment. Instead of playing the stimuli over headphones, participants were exposed to stimuli played over an array of loudspeakers, allowing for spatial separation of speech and the noise source. Results indicate that the ordering of the systems is the same in both the headphone and in the surround sound set-up, which confirms that experiments with headphones give a good prediction of performance in more realistic scenarios. Results from the surround sound set-up presented greater differences between the systems tested. Synthetic speech pre-

sented in a stereo set-up – headphones and stereo noise – was more intelligible than in a mono set-up – headphones but the same noise played in each ear. The mono and multichannel set-up results were very similar, even though the last scenario offered spatial separation, which the authors say indicates that the room impulse response has a negative impact on intelligibility.

Chapter 4

Evaluation of objective intelligibility measures

In this chapter, we present two experiments designed to evaluate objective measures of speech with regards to intelligibility prediction of HMM-generated synthetic speech in additive noise. We first explain how objective measures of speech operate and the different categories of measures. We then talk about how to use these measures to evaluate speech enhancement and speech modification algorithms and how to evaluate objective measures of intelligibility. We subsequently detail each of the experiments we designed, describing the listening set-up and speech material as well as presenting the subjective intelligibility scores and correlation coefficient results obtained in each one. Experiment I dealt mostly with non-modified synthetic speech and Experiment II with modified synthetic speech. We also comment on the impact of noise and modification type on the performance of the measures. All this is followed by discussions and conclusions. This work was partially published in Valentini-Botinhao et al. (2011b,a). Audio samples are available at `https://wiki.inf.ed.ac.uk/CSTR/Modifications`

4.1 Introduction

Objective measures for speech are tools for predicting dimensions such as quality and intelligibility, that otherwise would have to be obtained using subjective listening tests. Listening tests are time consuming, often expensive and not easily reproducible. Therefore objective measure are an attractive proposition. These measures can be used to evaluate and compare different methods of storing, transmitting, processing and generating speech, in the fields of speech coding and speech enhancement. In some

of these fields they have eventually replaced listening test experiments by providing reliable enough evidence to support improvements brought by the algorithms. But beyond their use for evaluation, could these measures be used as a control mechanism for speech generation or modification algorithms? The measure could, for instance, control the effect of a speech enhancement algorithm like noise suppression, dereverberation or source separation. Moreover a measure could be used to control the type and degree of speech modification required to enhance certain acoustic properties of clean speech, in the context of what we refer to here as speech intelligibility enhancement. If we wish to use objective measures in this way, we must first discover whether the predictions they make are accurate across different noise types, listening conditions, and speech modifications.

Several studies have shown the correlation between subjective and objective measures for quality and intelligibility prediction. Early studies concerned quality prediction for speech coders (Barnwell, T., III, 1980; Quackenbush et al., 1988; Kubichek et al., 1991). More recently-introduced measures and evaluation methods are designed to measure the quality and intelligibility of noise-corrupted speech processed using noise reduction algorithms (Hu and Loizou, 2006, 2008a; Ma et al., 2009; Taal et al., 2009; Ma and Loizou, 2011; Gomez et al., 2011) and dereverberation algorithms (Kokkinakis and Loizou, 2011).

We have seen in Chapter 2 that human speech production is affected by background noise, and that some of the modifications made by talkers are helpful to listeners. It is a natural step to incorporate similar behaviour into a text-to-speech (TTS) system. The TTS system would then be able to generate speech that is as intelligible as possible for any given listening condition. For this class of speech intelligibility improvement algorithms, where the clean speech is available and can be modified *before* being mixed with noise, there have been far fewer evaluation studies. Tang and Cooke (2011) report how well a set of objective measures perform for a range of modifications made to natural speech. However no extensive study has yet been performed involving many types of objective measures, diverse noise conditions and, crucially, with modified synthetic speech. But prior to conducting that study, we must first evaluate how well these objective measures correlate with subjective scores for unmodified synthetic speech. This questioning arises from the fact that these measures were originally proposed to predict quality or intelligibility of distorted natural speech. Here we are using them to evaluate distorted synthetic speech where distortion refers to the additive noise only and not the process of Text-To-Speech generation. The predictive power of the measures

Acronym	Measure
GP	Glimpse proportion (Cooke, 2006)
Dau	Dau measure (Christiansen et al., 2010)
STOI	Short Term Objective Measure (Taal et al., 2010)
SII	Speech Intelligibility Index (ANSI, 1997)
PESQ	Perceptual Evaluation of Speech Quality (Rix et al., 2001)
FWS	Frequency Weighted SNR (Tribolet et al., 1978)
WSS	Weighted Spectral Slope (Klatt, 1982)
CEP	Cepstral distance (Gray and Markel, 1976)
LSD	Log Spectral Distance (Gray and Markel, 1976)
IS	Itakura Saito distance (Gray and Markel, 1976)
LLR	Log Likelihood Ratio (Gray and Markel, 1976)

Table 4.1: Objective measures evaluated in this chapter.

could decrease as we are not using natural speech as a reference for clean undistorted speech.

To be able to produce acoustic changes leading to speech intelligibility improvement in a noisy environment, a statistical parametric speech synthesis system as described in Chapter 3 is a promising framework. This framework offers a great deal of flexibility during both model training and speech generation. There is the potential to generate synthetic speech that is most intelligible for a particular noise type and SNR without the need to train models on matched training data, but rather by modifying the model parameters or the generated speech parameters. In this study, we first investigate whether measures designed to predict the intelligibility of natural speech can also predict the intelligibility of hidden Markov model (HMM) generated synthetic speech, in noise. Our second concern is whether the measures can predict the impact on intelligibility of changes made to synthetic speech at a speech parameter level. Last, but not least, we investigate whether the modifications we chose to apply actually have a significant positive effect on subjective intelligibility.

In the course of two experiments, we evaluate the eleven different objective measures presented in Table 4.1. The relationship between speech quality and intelligibility is not a simple one: several well known speech enhancement – noise suppression – algorithms are not able to improve intelligibility although speech quality improves (Hu and Loizou, 2007a,b). This motivates us to include not only measures for in-

telligibility but also measures for quality, similar to the evaluation described in Taal et al. (2009). Four out of the eleven measures we evaluate are specifically designed to predict intelligibility – the Dau measure, the Glimpse proportion, the Short Time Objective Intelligibility (STOI) measure and the Speech Intelligibility Index (SII) – and one of them was specifically designed to measure quality – Perceptual Evaluation of Speech Quality (PESQ). The remainder are simpler measures, also commonly employed to measure quality.

In our first experiment, we evaluate the measures for intelligibility prediction of synthetic speech either unmodified, or modified with an ideal binary mask; the mask is of the type employed by some noise reduction algorithms. Some of the results of this experiment were previously reported in Valentini-Botinhao et al. (2011b). The results reported in this chapter were obtained from subjective and objective scores averaged across sentences *and* listeners, whereas in Valentini-Botinhao et al. (2011b), we averaged only across listeners. For this reason results presented here differ from the ones presented in that paper.

In our second experiment, we evaluate the same measures but this time for synthetic speech which has been modified by a range of simple, one dimensional, frame-wise modifications inspired by the acoustic properties of Lombard speech (Summers et al., 1988; Junqua, 1993; Castellanos et al., 1996). The modifications are: enhancement of spectral peaks, changes in fundamental frequency (F_0), shift of line spectral pairs (LSPs) and change in speaking rate. Some of the results from this experiment were previously reported in Valentini-Botinhao et al. (2011a). For both experiments, we also evaluate the effect of the modifications on subjective intelligibility. The results reported here for the Dau measure are slightly but not significantly different from the ones reported in the published papers as we found an error in the calculation of the measure further on in our work.

The remainder of this chapter is organised as follows. In Section 4.2, we describe the objective measures we chose to evaluate. In Section 4.3, we show how to quantify the measures' predictive power. In Sections 4.4 and 4.5, we present the evaluation results using data from experiments I and II respectively and in Section 4.6, we discuss our findings, followed by conclusions in Section 4.7.

4.2 Objective intelligibiliy measures

Several objective measures of quality or intelligibility of natural speech have been proposed, operating in a variety of manners by prioritising certain dimensions of the speech signal that are thought to reflect the perceptual cues that humans attend to when evaluating quality or intelligibility.

Predicting quality using objective measures has seen more success than predicting intelligibility. One of the most commonly used objective measures for speech quality, the Perceptual Evaluation of Speech Quality (PESQ) shows a high correlation with Mean Opinion Score (MOS) for various types of distortions (Rix et al., 2001). Objective measures of intelligibility do not correlate as well with subjective intelligibility scores.

The relationship between speech quality and intelligibility is not entirely clear. There have been various studies evaluating the usefulness of speech quality measures as predictors of intelligibility. One of the most recent (Taal et al., 2009) compared conventional methods based on SNR and linear prediction coefficients to perceptually-based measures and concluded that the latter are better predictors.

Fig. 4.1 gives two block diagrams representing the different ways these measures can operate, annotated with the names of the measures, which we will now introduce. The measures in the lefthand diagram in Fig. 4.1 are designed to estimate the intelligibility of speech with additive noise only as they require access to both the clean speech signal and the noise signal. They can be seen as audibility-based measures as they make predictions bases on the audibility of speech in noise. In this group there are SNR-based measures including Frequency Weighted Segmental SNR (FWS) (Tribolet et al., 1978) and the Speech Intelligibility Index (SII) (ANSI, 1997); there is also the auditory model-based measure called the Glimpse Proportion (GP) (Cooke, 2006).

The measures in the other group, on the righthand side of Fig. 4.1, can in principle calculate the intelligibility of other types of distortions, such as reverberation, coding and other non linear channel distortions like speech enhancement methods. To predict the intelligibility of a given speech signal, these measures need a reference undistorted speech signal and the distorted signal. In this group we find most of the measures evaluated in our current work, including spectral envelope-based distance measures like the Cepstral Distance (CEP), Log Spectral Distance (LSD), Itakura-Saito (IS) and Log-Likelihood Ratio (LLR) (Gray and Markel, 1976); there is also the Weighted-Spectral Slope Metric (WSS) (Klatt, 1982), the quality standard Perceptual Evaluation

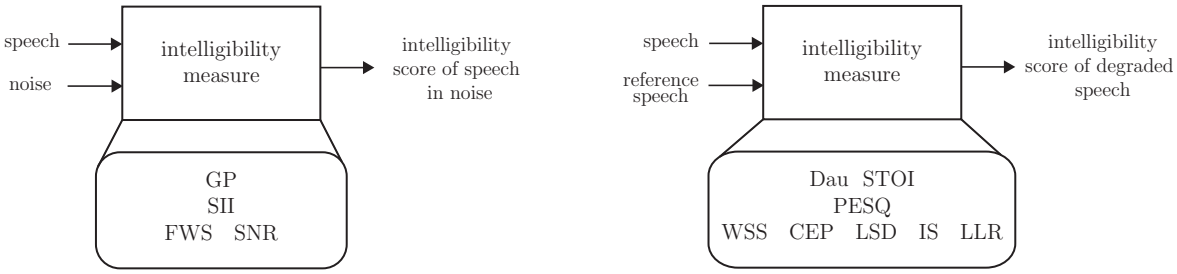


Figure 4.1: Objective measures of speech intelligibility and quality categories: measures for additive noise (left) and measures for any type of distortion (right). The acronyms refer to the measures we chose to evaluate in this study, with the exception of the SNR measure.

of Speech Quality (PESQ) (Rix et al., 2001), the Dau measure (DAU) (Christiansen et al., 2010) and the Short-Time Objective Intelligibility (STOI) measure (Taal et al., 2010).

All measures shown in Fig. 4.1 are called intrusive measures (based on so-called single-ended or full-reference models) because they need the clean or undistorted speech signal to make their prediction. Non-intrusive measures aim to predict a subjective dimension by analysing only the distorted speech. These types of measures are useful in applications where the clean undistorted speech signal is not available, like for instance quality monitoring during live calls (ITU, 2004) or quality and intelligibility prediction of disordered speech (Falk et al., 2011) and TTS systems (Falk and Möller, 2008; Falk et al., 2008; Norrenbrock et al., 2012). Here however we evaluate predictions of synthetic speech in noise intelligibility using only intrusive measures by using the synthetic speech as the clean undistorted signal.

In the following subsections, we describe how each of the eleven measures operate, grouped into four categories: spectrum-based, perceptually-motivated, standards and perceptual-model based.

4.2.1 Spectrum-based measures

The earliest approaches to quality prediction were based on simple metrics calculated on a spectral envelope derived from linear predictive analysis. In this group we have, amongst others, the Cepstral Distance (CEP), Log Spectral Distance (LSD), Itakura-Saito (IS) and Log-Likelihood Ratio (LLR) (Gray and Markel, 1976). These measures are usually calculated frame-by-frame and the final score is the average across frames, excluding silent frames. Here for simplicity we present the formulae for a single time

frame.

The Cepstral Distance (CEP) (Gray and Markel, 1976) is the unweighted Euclidian distance of cepstral coefficients:

$$d_{CEP} = \sqrt{\sum_{m=0}^{M} \left(c(m) - \overline{c}(m)\right)^2} \tag{4.1}$$

where $c(m)$ and $\overline{c}(m)$ are the cepstral coefficient m of the reference speech signal and of the speech signal being tested respectively.

The Log Spectral Distance (LSD) (Gray and Markel, 1976), as the name states, is the difference between the log spectrum envelopes:

$$d_{LSD} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\log\left(\overline{H}(\omega)\right) - \log\left(H(\omega)\right)\right]^2 d\omega \tag{4.2}$$

where $H(\omega)$ and $\overline{H}(\omega)$ are the spectral envelopes of the reference speech signal and of the speech signal being tested respectively.

The CEP and the LSD measures are related to each other. As seen in Chapter 3, the cepstral coefficients are the Fourier series representation of the logarithm of the spectral envelope. This means that the CEP measure converges to the LSD measure if the number of cepstral coefficients is made infinite (Gray and Markel, 1976).

The Log-likelihood ratio (LLR), also called the Itakura distance, is defined as:

$$d_{LLR} = \log_{10}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{H(\omega)}{\overline{H}(\omega)}\right|^2 d\omega\right) \tag{4.3}$$

$$= \log_{10}\left(1 + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{H(\omega) - \overline{H}(\omega)}{\overline{H}(\omega)}\right|^2 d\omega\right) \tag{4.4}$$

The LLR represents the log of the ratio of the energy of the residual signal of the reference and processed speech signal spectrum envelope. The residual signals are calculated by filtering the reference speech signal with the inverse of the spectrum envelope calculated from the reference and the processed signal, respectively. This means that the ratio is always larger than one as the residual signal energy will always be higher when speech and spectrum envelope are mismatched. Itakura (1975b) has shown that the ratio is actually a likelihood ratio under certain assumptions (Gray and Markel, 1976).

The Itakura-Saito (IS) distance proposed in Itakura and Saito (1970) is defined as:

$$d_{IS} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\left|\frac{H(\omega)}{\overline{H}(\omega)}\right|^2 - \log\left|\frac{H(\omega)}{\overline{H}(\omega)}\right|^2 - 1\right) d\omega \tag{4.5}$$

The IS distance is derived from Itakura's studies on the maximum likelihood formulation of the linear prediction coefficient calculation: speech is assumed to come from a Gaussian process, white noise filtered through an all-pole filter.

4.2.2 Perceptually-motivated measures

Following on from the methods described above are the measures that include some sort of explicit frequency-dependent weight, inspired by known properties of the human auditory system or psychoacoustics. In this work, we evaluate the Frequency Weighted Segmental SNR (FWS) (Tribolet et al., 1978) and the Weighted-Spectral Slope Metric (WSS) (Klatt, 1982).

The Frequency Weighted Segmental SNR (FWS) (Tribolet et al., 1978) is defined in a certain time frame as:

$$d_{FWS} = \frac{\sum_{k=1}^{K} w_k \, \mathrm{SNR}(k)}{\sum_{k=1}^{K} w_k} \tag{4.6}$$

where K is the number of frequency bands, w_k is a dynamic weight for frequency band k and:

$$\mathrm{SNR}(k) = 10 \log_{10} \frac{X^2(k)}{(X(k) - \overline{X}(k))^2} \tag{4.7}$$

is the SNR value at frequency band k, $X(k)$ and $\overline{X}(k)$ are the values of the filterbank amplitudes for the kth frequency band for the reference speech signal and the signal being tested. The weight w_k was originally defined as $|X(k)|^{0.2}$, which means higher weights are given to areas where the magnitude spectrum of the reference speech signal is higher.

The Weighted-Spectral Slope Metric (WSS) (Klatt, 1982) is defined for a single time frame as:

$$d_{WSS} = \sum_{k=1}^{K} w_k (S(k) - \overline{S}(k))^2 \tag{4.8}$$

where $S(k)$ and $\overline{S}(k)$ are the slopes for the reference speech signal and the signal to be tested. The slopes $S(k)$ are calculated as the first order differences in the critical-band spectra $X(k)$:

$$S(k) = X(k+1) - X(k) \tag{4.9}$$

$$\overline{S}(k) = \overline{X}(k+1) - \overline{X}(k) \tag{4.10}$$

and the weights w_k are chosen so that differences in slope around the spectral peaks are more important than around valleys and that the largest peak is more important

than others. This measure is more sensitive to spectral peak location than to their magnitudes, which makes it more appropriate for speech quality prediction than the spectrum-based measures described in the previous section (Klatt, 1982). Together with PESQ, the WSS measure is widely used for evaluating blind source separation and dereverberation algorithms (Di Persia et al., 2007, 2008).

4.2.3 Standards for quality and intelligibility

Various standards have been proposed to predict quality and intelligibility; these often incorporate some knowledge of psychoacoustics. The Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001) was designed as a measure for predicting the quality of speech signals transmitted over telephone lines and it became an ITU standard for evaluating telecommunication networks in 2000. This measure consists of the following stages: pre-processing, time alignment, auditory transform and disturbance processing. Pre-processing consists of gain equalization (based on the region between 350-3250 kHz - PESQ requires signals to be sampled at 16 kHz) and processing with a telephone headset frequency response filter. Time alignment is done through cross-correlation of temporal envelopes of the reference and degraded signal. The auditory transform calculates the loudness spectra of both signals and the disturbance is the difference between these contours. The measure is the disturbance signal averaged over time and frequency, mapped into a one to five scale to match the mean opinion score scale.

The Speech Intelligibility Index (SII) (ANSI, 1997) calculates a weighted SNR in the frequency domain, considering frequency-domain masking effects and auditory thresholds. The weights used in the SII calculation are fixed over time. It became a ANSI standard for intelligibility prediction of speech in additive noise in 1997. Because the SII is an SNR-based measure it cannot predict intelligibility in the presence of non-linear distortions; an extension for non-linear distortions has been proposed under the name cSII (coherence SII) (Kates and Arehart, 2005). The original SII measure is calculated over the entire signal. The extended SII (Rhebergen et al., 2006) is calculated over smaller intervals of the signal and then averaged across them, so it is then able to predict the intelligibility of speech in fluctuating noise. What we refer in our experiments as SII is in fact this extended version.

4.2.4 Perceptual model-based measures

The objective measures of intelligibility that have been shown to best correlate with subjective scores of natural speech intelligibility tend to be the ones that include elaborate auditory processing stages (Taal et al., 2009). These measures compare an internal representation of the clean reference speech signal with an internal representation of the noisy signal, or of the noise alone, in order to predict how intelligible the noisy signal is.

In this group, notable measures include the Dau measure (DAU) (Christiansen et al., 2010), based on the Dau model (Dau et al., 1996) of the effective processing which takes place in the human auditory system. The model gives a time-domain representation that incorporates aspects of temporal adaptation. It consists of: basilar membrane filtering (gammatone filterbank spaced on the equivalent rectangular bandwidth (ERB) scale (Moore and Glasberg, 1996)), hair cell transformation (half-wave rectification and low pass filtering), auditory nerve response (nonlinear adaptation loops with linear sensitivity to fast temporal changes and logarithmic steady-state response; the nonlinear loops can also simulate forward masking as output does not return to the initial condition immediately after stimuli switch off) and a modulation low pass filtering stage (the Dau model proposes a modulation filter bank but the measure simplifies this stage). The measure is effectively the weighted normalized correlation coefficient of the internal representation derived by the Dau model for the reference and the noisy signals. Different weights are given to time frames of different RMS levels (low-level, mid-level and high-level), as proposed by Christiansen et al. (2010).

The Glimpse proportion measure (GP) (Cooke, 2006) is derived from the Glimpse model for auditory processing. The measure is the proportion of spectral-temporal regions where speech is more energetic than noise, based on the idea that humans mainly attend to those 'glimpses' of speech that are not masked by noise. The spectral-temporal representation is obtained by the following stages: Gammatone filterbank whose central frequencies are linearly spaced on the ERB scale, temporal envelope extraction (absolute value operation and low pass filtering) and then averaging across limited time intervals.

The Short-Time Objective Intelligibility (STOI) (Taal et al., 2010) is the linear correlation coefficient between a time-frequency (T-F) representation of clean and a normalized T-F representation of noisy speech averaged over time frames. The T-F representation is obtained by: one-third octave band analysis of windowed time frames

of 25.6 ms with 50 % overlap. The normalized T-F representation sets the energy of the noisy T-F representation to be the same as the clean speech local energy which depends on the previous 30 time frames, around 400 ms. The normalization also includes a clipping stage to set a upper bound for signal to noise distortion. This measure is claimed to work especially well for conditions where noisy speech is processed by a T-F weighting algorithm for noise reduction or speech separation (Taal et al., 2010).

4.3 Prediction of the intelligibility of modified speech in noise

In order to evaluate an objective measure we need to compare intelligibility scores predicted by the objective measure with subjective intelligibility scores obtained from listening tests using the same speech material. The material should cover a wide range of listening conditions, i.e. different noise types and noise levels, so we can draw conclusions about the stability and generality of the measure's performance.

A commonly used evaluation metric is the normalized correlation coefficient – Pearson's correlation – between the objective and subjective scores. The higher the correlation, the better the measure performs. The correlation coefficient is defined as:

$$r = \frac{\sum_{n=1}^{N}(S_n - \bar{S})(M_n - \bar{M})}{\sqrt{\sum_{n=1}^{N}(S_n - \bar{S})^2 \sum_{n=1}^{N}(M_n - \bar{M})^2}} \tag{4.11}$$

where S_n is the subjective score for listening condition n, $\bar{S}$ is the average score obtained for all conditions in that group, M_n is the objective score given by the measure for listening condition n, $\bar{M}$ is the average score given by the measure for all conditions in that group. A listening condition can refer to the noise type, SNR and speech (unmodified or enhanced).

Another metric that is usually used to evaluate objective measures (Hu and Loizou, 2008a; Ma and Loizou, 2011) is the estimate of the standard deviation of the error of using an objective measure rather than subjective scores. We refer to this value as the standard deviation of the error:

$$\sigma_e = \sigma_s \sqrt{1 - r^2} \tag{4.12}$$

where σ_s is the standard deviation of the subjective intelligibility scores in a given condition. The standard deviation of the error will be smaller for measures that are

better intelligibility predictors. Although the scatter plots presented in this chapter show the subjective scores as percent scale 0-100, we present tables with values of σ_e on scale of 0-1 (instead of 0-100) as is conventional for objective measure evaluation (Hu and Loizou, 2008b) (Ma and Loizou, 2011).

Distance-type measures like CEP, LSD, IS, LLR and WSS *increase* when speech is less intelligible, whereas correlation- and audibility-based measures like FWS, PESQ, SII, Dau, STOI and GP *decrease*. The objective measures are also not necessarily linearly correlated with subjective scores. It is therefore common to apply a mapping before calculating the correlation coefficient. This will take care of the different dynamic ranges in the various measures and the non linear relationship. In this chapter, we show correlation coefficients obtained after applying a logistic mapping to the objective data. We found the parameters of the logistic function for unmodified and modified speech signals separately using all the data available as we have very few data points, as has also been done in other objective measure evaluations (Taal et al., 2009). Other studies have divided the data into $2/3$ for obtaining the mapping and $1/3$ for obtaining the correlation coefficient (Christiansen et al., 2010). Since we are testing fewer conditions (noise types, SNRs and voice types) we do not have enough data points for making such a partition. Therefore, as seen in other evaluation studies (Ma et al., 2009), we also provide correlation results when not performing any mapping: see Appendix A.

We are interested in evaluating these measures in the context of speech intelligibility enhancement, that is modifying clean speech so that the mixture of speech in noise is more intelligible. Fig. 4.2 shows how we can use these objective measures when clean speech is modified. When a measure requires access to the speech and noise separately (measures on the left-hand diagram of Fig 4.1), the evaluation is straightforward: the speech signal is the modified speech (top diagram of Fig. 4.2). However when the measure needs both the corrupted speech and a clean reference speech signal (measures on the right-hand diagram of Fig 4.1) , a choice must be made to which signal should be used as the reference. The corrupted speech will always be the modified speech plus noise, as this is the signal we are predicting upon (this is also the signal that participants hear in our listening tests). As far as the reference speech signal is concerned we see here two options: use clean modified speech as the reference (middle diagram of Fig. 4.2) or clean unmodified speech (bottom diagram of Fig. 4.2). The first option judges only the impact of the additive noise; the second, judges the impact of both the modification and of adding noise. Depending on the type of modification

Measure	CPU time (secs.)	Elapsed time (secs.)
GP	1.14	0.94
Dau	1.35	1.34
STOI	0.55	0.56
SII	0.46	0.46
PESQ	0.32	0.31
FWS	0.24	0.24
WSS	1.08	1.08
CEP	0.06	0.06
LSD	0.41	0.41
IS	0.26	0.26
LLR	0.26	0.26

Table 4.2: Average processing time to analyse one sentence using Matlab. All measures were calculated using a fixed window length of 30 ms with a 10 ms time shift.

and measure, one choice might be better than the other, which is something that we are also investigating here.

For the experiments shown here, we used the implementation found in (Loizou, 2007) of the spectrum-based and perceptually-based measures, and of the PESQ standard. We used the extended version of the SII measure. The implementations of STOI and GP were each provided by the respective author of the proposed measure and the Dau model was obtained from the computational auditory signal processing and perception (CASP) model (Jepsen et al., 2008). All measures are calculated over a 30 ms window frame, with a frame shift of 10 ms. Table 4.2 shows the processing time that each measure takes to predict the intelligibility of a sentence used in this experiment. We can see that measures like Dau, GP and WSS have a substantial computational cost when compared to other measures – almost twice as slow to compute. The measures with lower computational cost are: FWS, IS and LLR.

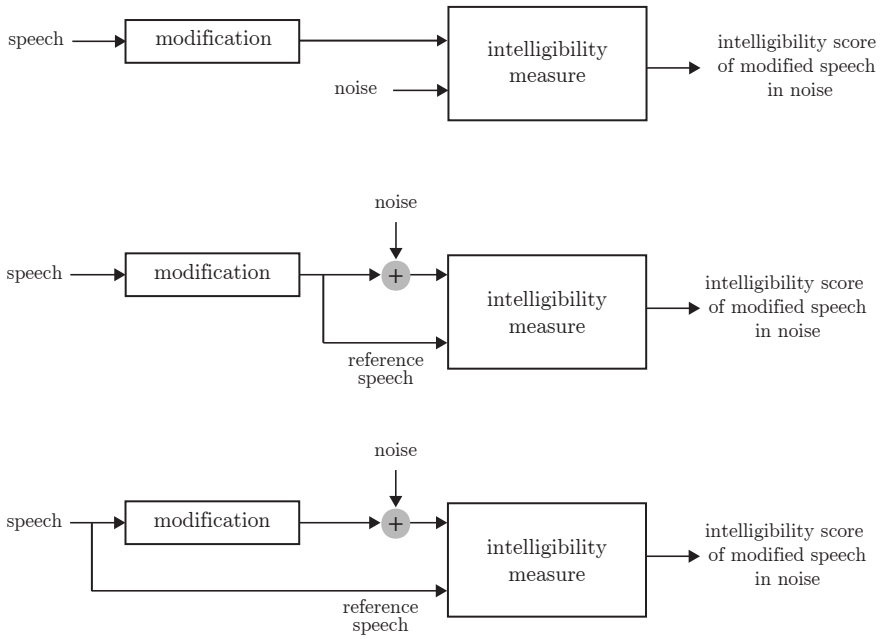


Figure 4.2: Block diagram describing how to use an objective measure to evaluate a speech modification algorithm for additive noise: objective measures that do not require reference speech signal (top) and objective measures that require reference speech signal (middle and bottom). The middle and bottom diagram reflect different choices for reference speech signal, modified and unmodified speech respectively.

4.4 Experiment I: synthetic speech and modifications based on the ideal binary mask

4.4.1 Experimental data

Similar to the evaluation described in Taal et al. (2009), we used so-called matrix sentences of the form "name verb numeral adjective noun". Each word in the sentence is chosen from an English ten-word list found in Dreschler (2006). In total, 108 sentences were synthesized using a statistical parametric synthesizer toolkit (HTS) (Tokuda et al., 2009). The synthesis models were trained with 4000 sentences from a professional male British English speaker named *rjs*.

The acoustic model we used for synthesizing speech was a hidden semi Markov model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. We used one stream for the spectrum and three streams for the logF_0. The models used 45 dimension mel-generalized cepstrum line

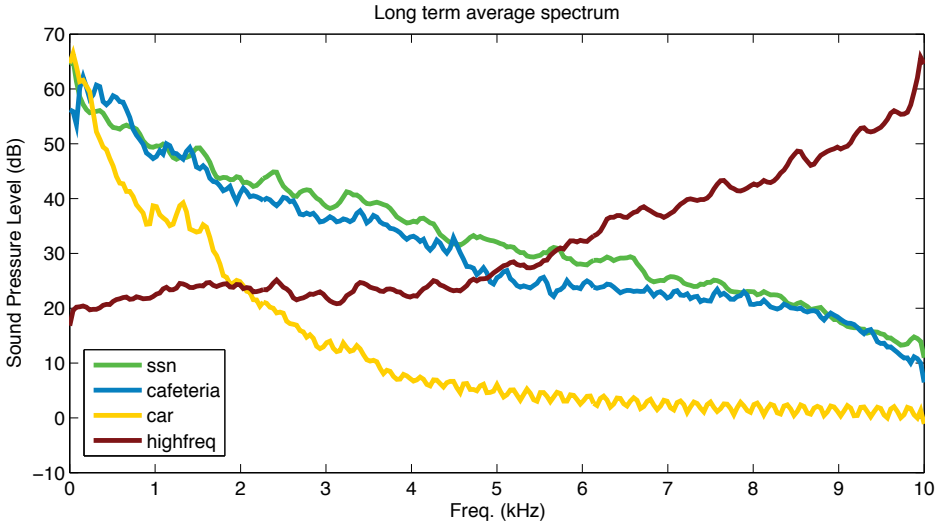


Figure 4.3: Long Term Average Spectrum (LTAS) in sound pressure level of the noises used in Experiment I and II: speech-shaped, cafeteria, car and high frequency noise. The speech-shaped noise LTAS was set to match the cafeteria LTAS.

spectral pairs (MGC-LSP) acoustic features as spectral features (Tokuda et al., 1994). For the excitation parameters we extracted F_0 and 25 aperiodicity band aperiodicity energies to construct the mixed multi-band excitation signal (Kawahara et al., 2001). The training sentences were sampled at 48 kHz. The synthesized speech was produced at 48 kHz then downsampled to 20 kHz. This downsampling was necessary as the STOI measure operates only at this lower sampling rate. For PESQ it was necessary to downsample to 16 kHz.

We used four different types of noise: speech-shaped, cafeteria, car and high frequency noise. The Long Term Average Spectrum (LTAS) of the utilized noises can be seen in Fig. 4.3. The LTAS is calculated as the power spectral density averaged across time frames of 10 ms length and 50 % overlap. This averaged representation is then presented in dB sound pressure level. The speech-shaped and high frequency noises were generated from filtered white noise. The cafeteria and car noises were actual recordings and are non stationary. The car noise has a periodic content that changes with time. The high frequency noise is not a realistic signal but is used here in order to investigate whether Lombard inspired modifications – usually caused by the presence of a low frequency masker – could bring intelligibility benefits in the high frequency noise. All noises were added at five different SNRs, chosen to be: -10 dB, -5 dB, 0 dB, 5 dB and 10 dB for speech-shaped noise and cafeteria, -30 dB, -25 dB, -20 dB, -15 dB and -10 dB for car noise and -40 dB, -35 dB, -30 dB, -25 dB and

-20 dB for high frequency noise.

In total we created 36 different listening conditions, by varying the noise and speech modification. The first set of conditions, where no modification was applied to the speech, constitutes 20 of these (four different additive noises each added at five different SNRs). The second set employed modified speech and constitute the other 16 conditions (four noises added at two different levels to speech modified with two different modification strengths).

4.4.2 Speech modification

In this experiment, the modified speech was created from clean speech by applying an Ideal Binary Mask (IBM) to it (Brungart et al., 2006; Kjems et al., 2009). The IBM is defined as a time-frequency (T-F) binary filter, with T-F bins equals to '1' when the local SNR is above a certain threshold and '0' otherwise. The T-F space is an auditory inspired one as it is achieved by a frequency decomposition that uses a Gammatone filterbank whose center frequencies are linearly spaced on the equivalent rectangular bandwidth scale. In the context of noise suppression, the IBM is applied to the noisy mixture and it provides the only known criterion that can significantly improve speech intelligibility of a noisy mixture signal (Loizou and Kim, 2011), for both normal-hearing (Brungart et al., 2006; Kjems et al., 2009) and hearing-impaired listeners (Hu and Loizou, 2008b). However it requires the noise signal to be separate from the speech signal, something that is not available to noise suppression algorithms.

In our experiments, the mask is calculated according to the noisy mixture but applied to the clean speech before mixing it with noise. Since the root mean square energy of clean speech is normalized to remain unmodified, what this process does is to concentrate the signal in those time-frequency bins of speech that will be more energetic than the noise while removing the bins that will be below the level of the noise, with the aim of increasing the intelligibility of the mixture. The local SNR threshold used to create the IBM was a parameter we varied, leading to the two different strengths of modification: $s_1 = \mathrm{SNR}_S + 5$ and $s_2 = \mathrm{SNR}_S + 10$, where SNR_S refers to the SNR at a sentence level. The strength s_2 is the strongest modification as the local SNR level required for a time-frequency bin to be left unprocessed is higher, meaning that speech is highly filtered in this condition. This modification was tested only at the lowest two SNRs (-10 dB and -5 dB). The authors in Kjems et al. (2009) provided us with an implementation of the IBM mask processing.

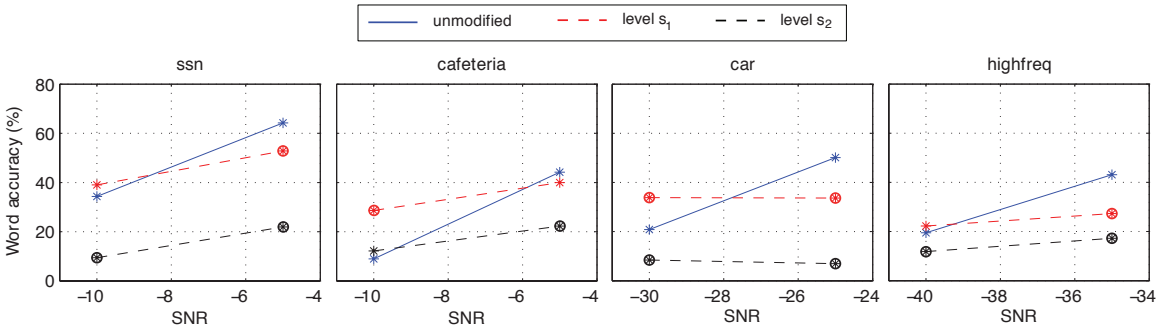


Figure 4.4: Experiment I: average subjective scores (in percent) of unmodified and modified synthetic speech. The applied modification was to filter the clean speech with an ideal binary mask. There were two SNR conditions for binary mask construction: s_1 and s_2. All scores significantly different to the one obtained with the unmodified speech are marked with circles.

4.4.3 Listening set-up

A total of 41 native English speakers aged mostly between 20 and 30 years old with no reported listening impairments participated in the listening test. Each participant listened to each condition three times with different sentences each time and in a random order. All signals were played at 20 kHz sampling rate over headphones to participants in sound-isolated booths. Each individual sentence could be played only once before the participant had to type in what he or she heard.

4.4.4 Subjective intelligibility scores

We calculated the subjective score of word accuracy rate (WAR) as the percent of correct words in a sentence (Hu and Loizou, 2007b), taking into account misspelling and spelling variations. Fig. 4.4 shows the average subjective scores in percent for the listening conditions where synthetic speech was processed by an ideal binary mask. The solid line connects the values obtained by unmodified speech and each dashed line connects values for modified speech at the different modification strength s_1 and s_2. The circles indicate significant differences at a 5 % level, when compared to the corresponding score for the unmodified speech. As we can see, for most noise types and SNRs, this particular modification decreases intelligibility. Only for the lowest SNRs there was a significant increase in intelligibility, in the presence of cafeteria noise (from 10 % to 28 %) and of car noise (from 21 % to 34 %). Only in this condition did applying an IBM with the chosen configuration have a positive impact on intelligibility.

	Dau	GP	STOI	PESQ	WSS	SII	FWS	IS	CEP	LSD	LLR
r	0.94	0.94	0.90	0.83	0.77	0.77	0.67	0.38	0.32	0.32	0.32
σ_e	0.10	0.10	0.13	0.16	0.18	0.18	0.21	0.27	0.27	0.27	0.27

Table 4.3: Experiment I: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech.

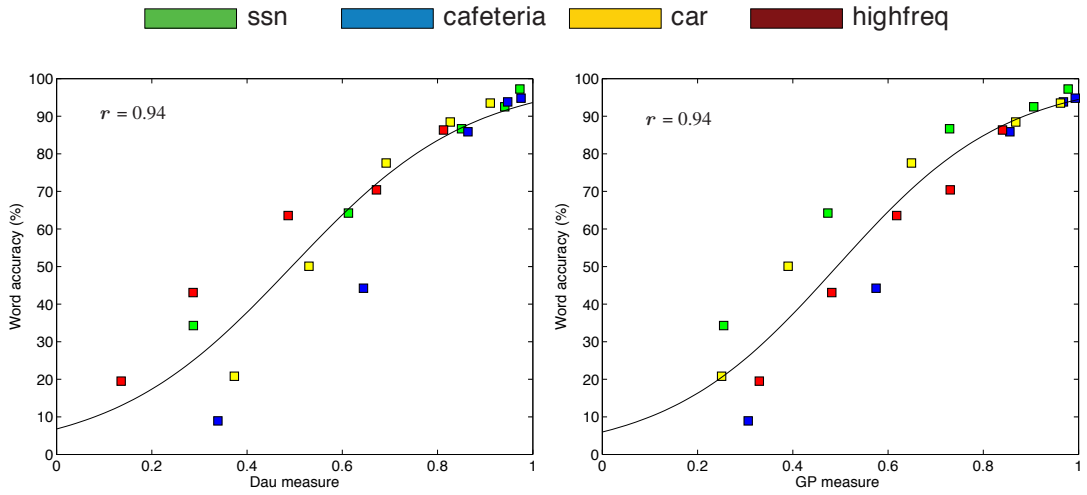


Figure 4.5: Experiment I: scatter plots of word accuracy (%) against the Dau measure (left) and GP measure (right) predictions of accuracy, for *unmodified* speech in all noise conditions. Each point refers to the score average across sentences and listeners for a certain noise condition (noise type and level).

4.4.5 Evaluation results

We calculated the normalized correlation coefficient and the standard deviation of the error using the subjective scores obtained in each listening condition, averaged across listeners and sentences.

Table 4.3 shows the correlation coefficients obtained for the unmodified speech material in decreasing order according to the best results each measure achieved. We can see that the Dau and Glimpse Proportion (GP) measures are the better predictors for intelligibility, with correlation coefficients of 0.94 both and standard deviation of the error of 0.10 both. The STOI measure is the third best measure with a correlation coefficient of 0.90 and a standard deviation of 0.13. The spectrum-based measures LSD, CEP, LLR and IS performed the worst: CEP, LSD and LLR achieved the same correlation coefficient of 0.32 and IS a higher coefficient of 0.38. The standards and perceptually-motivated ones are somewhere in between, PESQ obtaining the best result

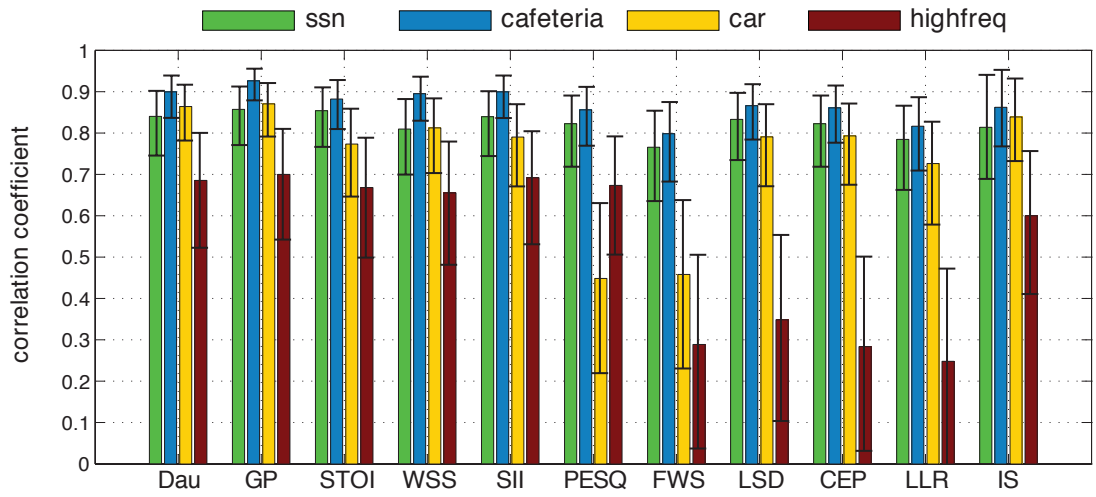


Figure 4.6: Experiment I: correlation coefficient and confidence intervals for *unmodified* synthetic speech, broken down by noise type. The measures are ordered in decreasing order of correlation coefficient averaged across noise type.

of 0.83 followed by WSS and SII with 0.77 and FWS with 0.67. When we compare the results in Table 4.3 with the correlation coefficients obtained across different natural speech evaluations (Taal et al., 2009; Cooke, 2006; Christiansen et al., 2010; Taal et al., 2010) we observe a loss of prediction performance for all measures when the speech is synthetic rather than natural, particularly for the SII measure and the spectrum-based measures. Fig. 4.5 shows scatter plots of subjective scores against objective scores obtained by the measures that had the best performance with unmodified synthetic speech: Dau and GP. Each point corresponds to a different condition of noise type and level (20 points in total).

To better understand the effect of the noise type on the objective measures, we calculated correlation coefficients for each noise type; these results are shown in Fig. 4.6 with their 5 % confidence intervals. We can see that all measures exhibit a drop in performance for the high frequency noise case, especially the spectrum-based measures which do not incorporate psychoacoustic knowledge about human auditory sensitivity to high frequency noise. For these results, as we did not have many data points for each noise type, we averaged the subjective scores across listeners, but not across sentences. We can see that many measures obtain correlation coefficients above 0.8 in all noise types, except the high frequency noise. The drop of performance under this noise might explain the overall drop in performance of measures like LSD, CEP, LLR and IS. Car noise seems to be challenging for the PESQ and FWS measures, although PESQ obtained quite a high correlation coefficient when considering all noises together. The

	Dau	GP	STOI	PESQ	WSS	SII	FWS	IS	CEP	LSD	LLR
Case 1											
r	0.01	0.52	0.42	0.30	0.12	0.37	0.07	0.22	0.04	0.13	0.14
σ_e	0.13	0.11	0.12	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.13
Case 2											
r	0.01	-	0.42	0.30	0.55	-	-	0.70	0.13	0.06	0.12
σ_e	0.13	-	0.12	0.13	0.11	-	-	0.09	0.13	0.13	0.13

Table 4.4: Experiment I: correlation coefficient r and standard deviation of the error σ_e for *modified* synthetic speech, when using modified speech (Case 1) or unmodified speech (Case 2) as the reference clean speech signal for calculating the objective measures that require a reference signal. The results of the measures that do not require a reference speech signal - GP, SII and FWS - are presented as belonging to Case 1.

measures that are more robust to different noise types are Dau, GP, STOI and WSS, particularly the first three measures as they showed good prediction performance when all noise types are considered.

We now present the results with modified speech in Table 4.4. We can see that all measures perform worse for *modified* synthetic speech. When the modified speech is taken to be the reference, Case 1 in the Table, the GP and the STOI measures seems to perform best, obtaining a correlation coefficient of 0.52 and 0.42 respectively and the smallest standard deviations 0.11, 0.12. This result could be expected because this measure predicts intelligibility from the proportion of time-frequency bins that are above the noise, which matches the type of modification we performed. Table 4.4 also shows the correlation coefficients obtained when the unmodified speech is used as the reference signal, Case 2. The correlation coefficient of the measures that do not need a reference signal (the GP, FWS and SII) are presented as Case 1 as they only use the modified speech signal. It seems that using the unmodified speech signal as the reference signal improves the correlation coefficient considerably for the IS and WSS measures, that now obtain correlations of 0.70 and 0.55 and standard deviations of the error of 0.09 and 0.11 respectively. The other measures seem to have similar performance on both cases. Fig. 4.7 shows the scatter plots of subjective scores against objective scores obtained by the IS and WSS measures with modified synthetic speech.

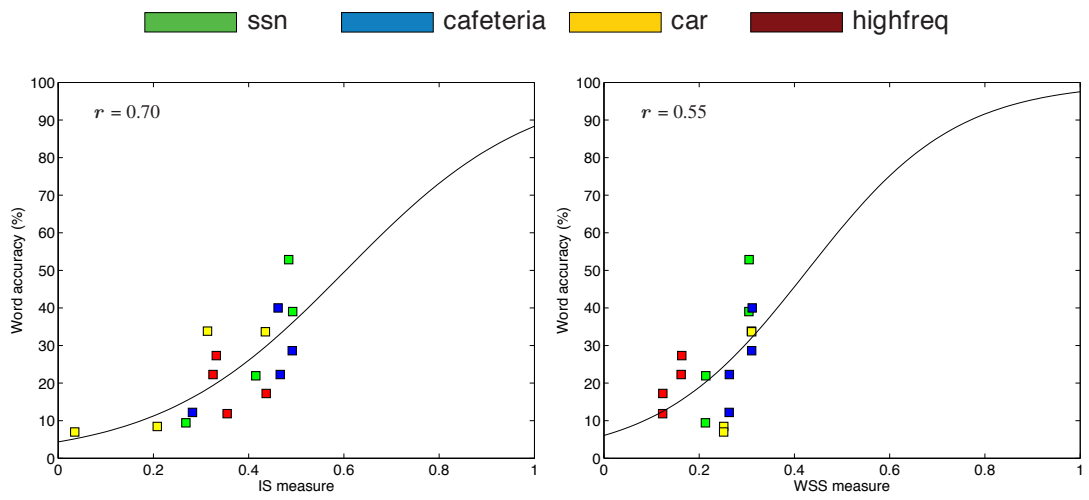


Figure 4.7: Experiment I: scatter plots of word accuracy (%) against the IS measure (left) and WSS measure (right) predictions for the *modified* speech. Each point represents scores averaged across sentences and listeners for a certain noise condition (noise type and level).

4.5 Experiment II: synthetic speech and modifications based on Lombard speech

4.5.1 Experimental data

In total, we synthesized 96 different sentences using the same models and configuration as used in Experiment I. The format of the test sentences was once again "name verb numeral adjective noun" (i.e., matrix sentences), with each word being chosen from a ten-word list taken from (Dreschler, 2006).

We used the same four noises as in Experiment I: speech-shaped noise (ssn), cafeteria, car and high frequency noise at four different SNRs. For this experiment, we performed a small calibration test to choose the SNRs that corresponded roughly to word accuracies of 20 %, 40 %, 60 % and 80 % for each noise type when using unmodified synthetic speech. For this calibration test we used 9 listeners, each heard sentences played at 24 different SNR conditions mixed with each four maskers. Fig. 4.8 shows the average subjective word accuracy obtained in the calibration test and the fitting curves calculated for each noise type. We can see that different SNR values can lead to similar intelligibility levels depending on the noise type, which is precisely why this calibration step is required before the main listening test. The high frequency noise requires a much lower SNR than speech-shaped and cafeteria noise to result in the same

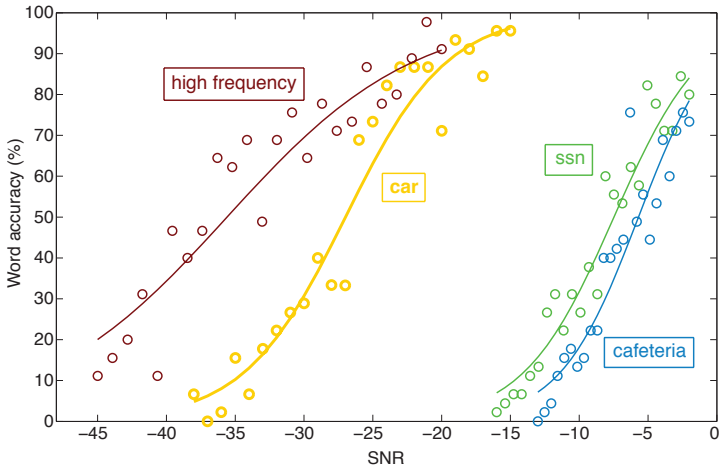


Figure 4.8: Experiment II - calibration test results: fitting curves and scatter plots of subjective word accuracy in percent against SNR for each noise type obtained in the calibration test with *unmodified* synthetic speech. Each point represents scores averaged across sentences and listeners for a certain noise condition (noise type and SNR).

	WAR (%)			
	20 %	40 %	60 %	80 %
ssn	-11.8	-8.8	-6.2	-3.1
cafeteria	-9.5	-6.8	-4.6	-1.9
car	-31.9	-28.4	-25.5	-22.0
high frequency	-43.5	-37.6	-32.7	-26.8

Table 4.5: Experiment II: SNR values (in dB) of each noise type corresponding to four different word accuracy rates in percentages.

intelligibility level. Note that the slope of the curve is also noise-dependent: small SNR differences can lead to larger changes in intelligibility for speech-shaped and cafeteria noise, compared to high frequency and car noise. The SNRs corresponding to the 20 %, 40 %, 60 % and 80 % WAR conditions are shown in Table 4.5.

4.5.2 Speech modifications

As mentioned in Chapter 2, it has been observed that, compared to speech produced in quiet conditions, speech produced in noise tends to present sharper spectral peaks, higher fundamental frequency, flatter spectral tilt and longer duration (Summers et al.,

1988; Junqua, 1993; Castellanos et al., 1996). The exact phenomena observed depend on the phonetic unit and the effect is slightly different for female and male speakers. In this experiment our goal is to discover which of these acoustic changes – when presented individually – actually contribute towards intelligibility increases, and whether their effect can be predicted by objective measures. In order to simulate the acoustic properties of natural Lombard speech we use one control parameter, referred here as strength, to individually modify the spectrum envelope, fundamental frequency and speech rate. The following modifications were chosen:

- **peak**: spectral peak enhancement performed using the post filter described in Ling et al. (2006) using two different levels for the parameter the authors refer to as α: s_1 (α=0.7) and s_2 (α=0.6). A lower value of α reflects stronger enhancement. The peak enhancement is done recursively by readjusting the line spectral pairs (LSPs) so that consecutive frequencies are closer together with decreasing α as follows:

$$\overline{\omega}_k = \omega_{k-1} + d_{k-1} + \frac{d_{k-1}^2}{d_{k-1}^2 + d_k^2}[(\omega_{k+1} - \omega_{k-1}) - (d_k + d_{k-1})] \tag{4.13}$$

  where ω_k is the k-th line spectral frequency and $d_k = \alpha\,(\omega_{k+1} - \omega_k)$.

- **F_0**: changes in the fundamental frequency (F_0): one reduction s_1 (30 % lower) and two increases s_2 and s_3 (30 % and 50 % higher), applied directly to the generated sequence of F_0.

- **LSP shift**: frequency shift of Line Spectral Pairs (LSPs) as described in (McLoughlin and Chance, 1997) at three different strengths s_1 ($\gamma = 1.015$), s_2 ($\gamma = 1.025$) and s_3 ($\gamma = 1.05$), always shifting towards the high frequency region. The shift is performed on the k-th line spectral frequency ω_k using the following transformation:

$$\overline{\omega}_k = \omega_k + \omega_k(\gamma - 1)(\pi - \omega_k)/\pi \tag{4.14}$$

  If $\gamma > 1$ the shift is towards the high frequency, if not the LSP are shifted towards the low frequency, with $\gamma = 1$ LSPs are not modified. This modification is applied to all frames and in the Mel scale.

- **rate**: changes in the speaking rate are obtained by modifying the parameter Yoshimura et al. (1998) call ρ, different values of ρ were set by changing the

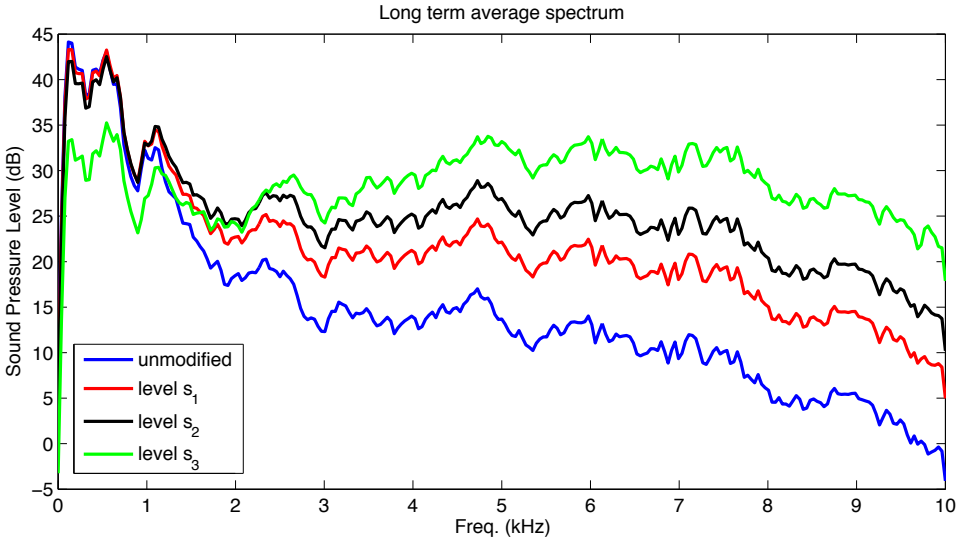


Figure 4.9: Experiment II: long term average spectrum (LTAS) of an unmodified utterance compared to modified speech in which the LSPs are shifted towards higher frequencies.

scale factor φ defined in Section 3.2.4. We increased the speaking rate in percentages of the original duration: s_1 (60%) and decreased it s_2 (140%) and s_3 (200%).

We chose the various values of modification strength such that they generated audible differences when compared to the clean speech condition. All modifications were applied to the generated sequence of speech parameters at a frame level obtained using previously-trained models and before they were passed to the synthesis filter. The modifications of speaking rate had a different impact on each phonetic unit, as the effect is proportional to the standard deviation of their duration distribution. This means that vowel duration tends to increase more compared to consonant duration, as observed in Lombard speech (Junqua, 1993).

To illustrate the effect of shifting the LSPs we plot the LTAS of unmodified speech and LSP-shifted speech in Fig. 4.9. We can see that the spectral tilt becomes flatter and we expect that the formant frequencies also increase. The average spectral tilt, computed as described in Lu and Cooke (2009b) of the unmodified speech was -1.51 dB per octave and for the three strengths of shifts, s_1, s_2 and s_3, the average spectral tilt was found to be -1.02 dB, -0.69 dB and -0.09 dB per octave, respectively.

In order to add the speech signals to noise at the same SNR, we first normalized both the unmodified and the modified speech signals sentence by sentence to yield the

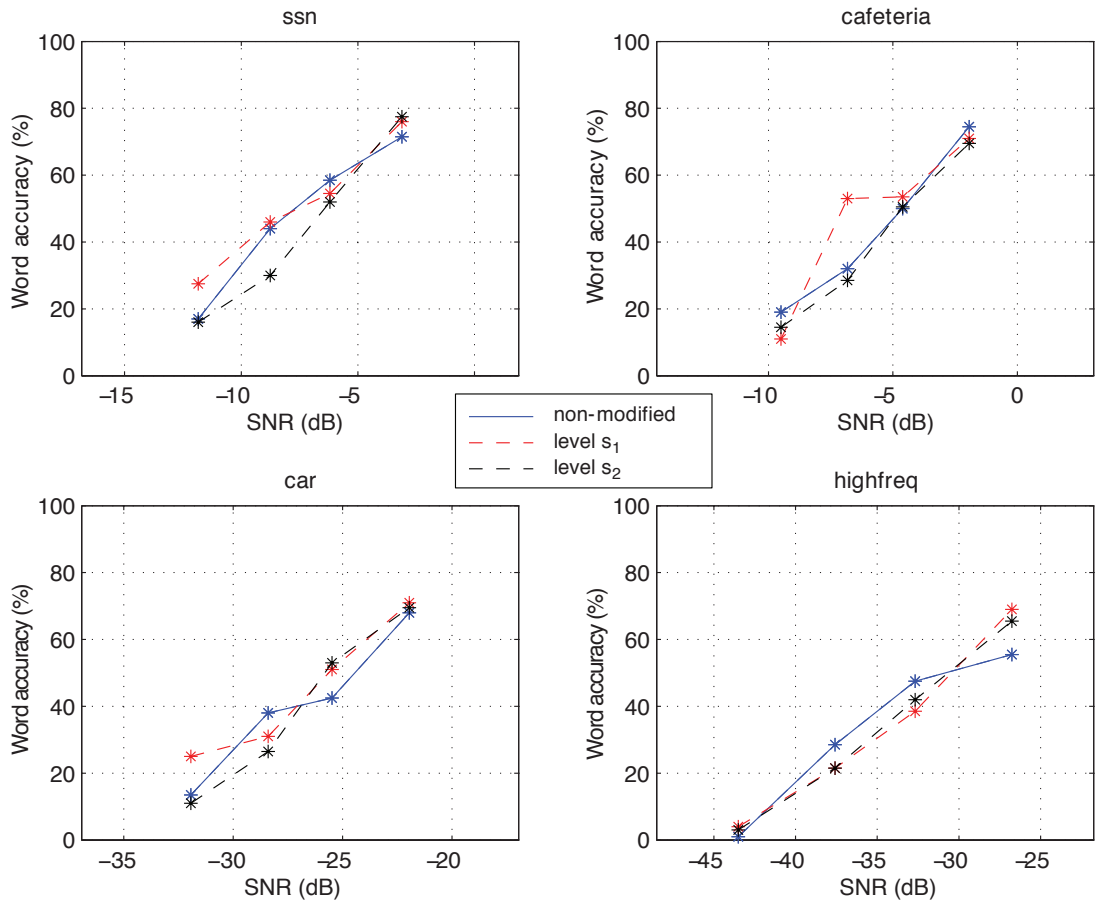


Figure 4.10: Experiment II: average subjective word accuracy for each noise type and SNR level. The curves represent different modification strengths for *spectral peak* enhancement. No significant differences were found among these scores.

same overall signal level. This means that any intelligibility change observed when applying a certain modification would not be the result of changes in overall energy levels. In total we generated 192 distinct listening conditions from all combinations of noise type (four), SNR (four), speech modification type (four) and modification strength (three), including an unmodified case for the spectral peak modification.

4.5.3 Listening set-up

We had a total of 88 native English speakers aged mostly between 20 and 30 years old and with no reported hearing impairment participating in the listening experiment. As we had so many listening conditions, each participant listened to only one quarter of all possible conditions twice, each time with different sentences - a total of 96 sentences. Across each group of 4 listeners, all conditions were covered. The order of sentences

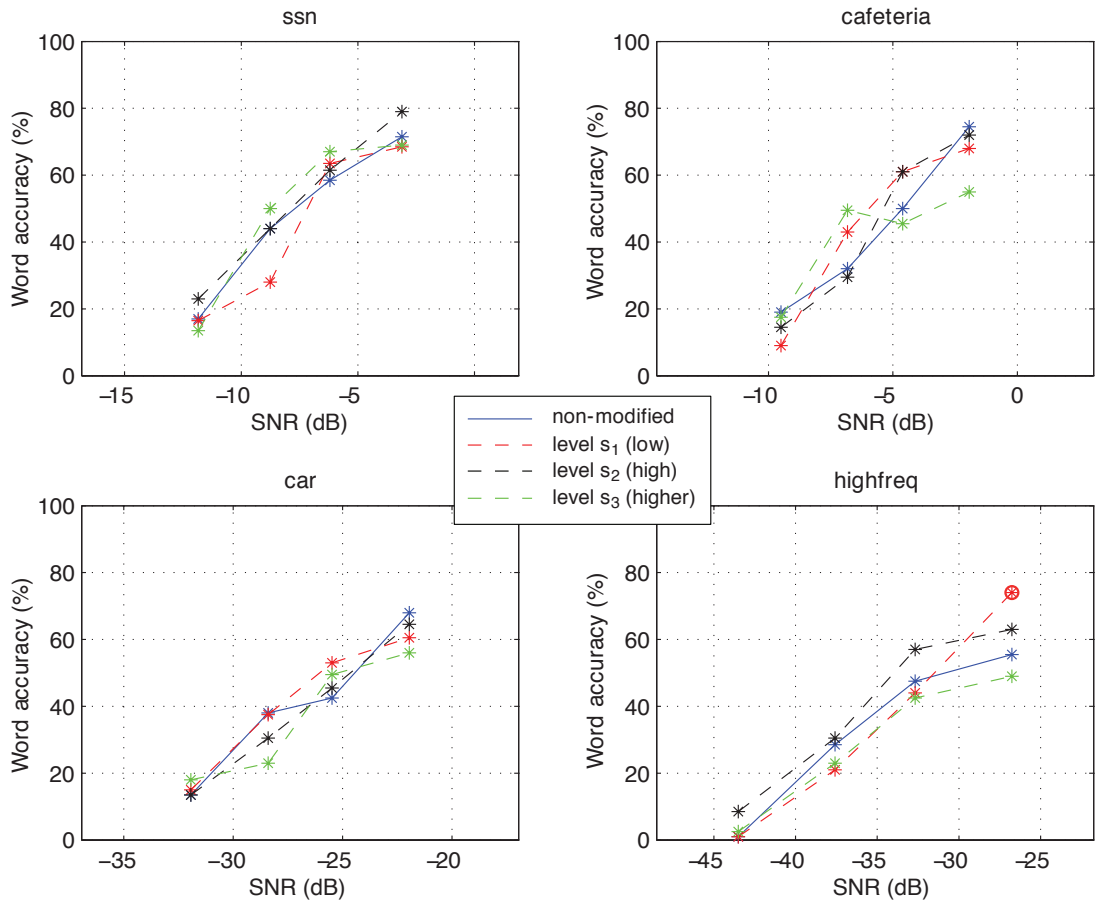


Figure 4.11: Experiment II: average subjective word accuracy for each noise type and SNR level. The curves represent different modification strengths for *fundamental frequency* changes. No significant differences were found among these scores.

and listening conditions was random, as was the selection of listening conditions.

We played all signals over headphones to participants in sound-isolated booths and each individual sentence could be played only once before the participant had to type in what he or she heard. Before the actual test took place each participant heard a few samples, including several listening conditions with modified and unmodified speech, to familiarize themselves with the task.

4.5.4 Subjective intelligibility scores

Following the same procedure adopted in Experiment I, we calculated the subjective word accuracy as the percent of correct words in a sentence, taking into account misspelling and spelling variations. The word accuracy results are displayed separately for each modification type in Figs. 4.10, 4.11, 4.12 and 4.13. These figures show the

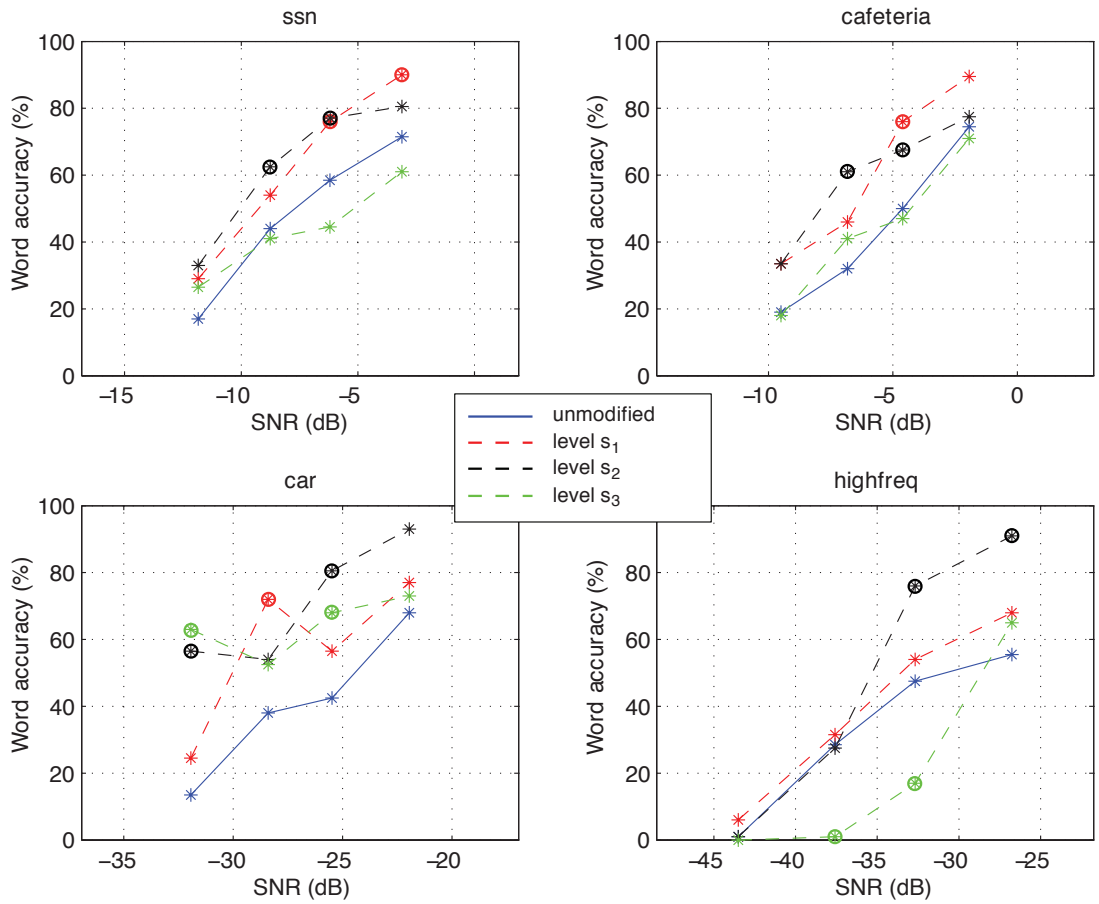


Figure 4.12: Experiment II: average subjective word accuracy for each noise type and SNR level. The curves represent different modification strengths for *LSP shift* changes. The scores significantly different to the score obtained with unmodified speech are marked with circles.

subjective scores, averaged over sentences and participants, for each noise type at each SNR value for these modification types. The solid line is a piecewise linear representation of the unmodified condition accuracy data point results and the dashed lines correspond to different modification strengths: s_1, s_2 and s_3. The circles indicate when a value is significantly different from the score for unmodified speech in the same noise condition at a 5 % level.

Although the authors of the post filter report an increase in speech quality (Ling et al., 2006), Fig. 4.10 shows that spectral peak enhancement did not have any significant impact on intelligibility in this experiment. Increasing the fundamental frequency, see Fig. 4.11, showed no significant impact on intelligibility scores either. Lowering F_0, strength s_1, provided a significant improvement for high frequency noise at the

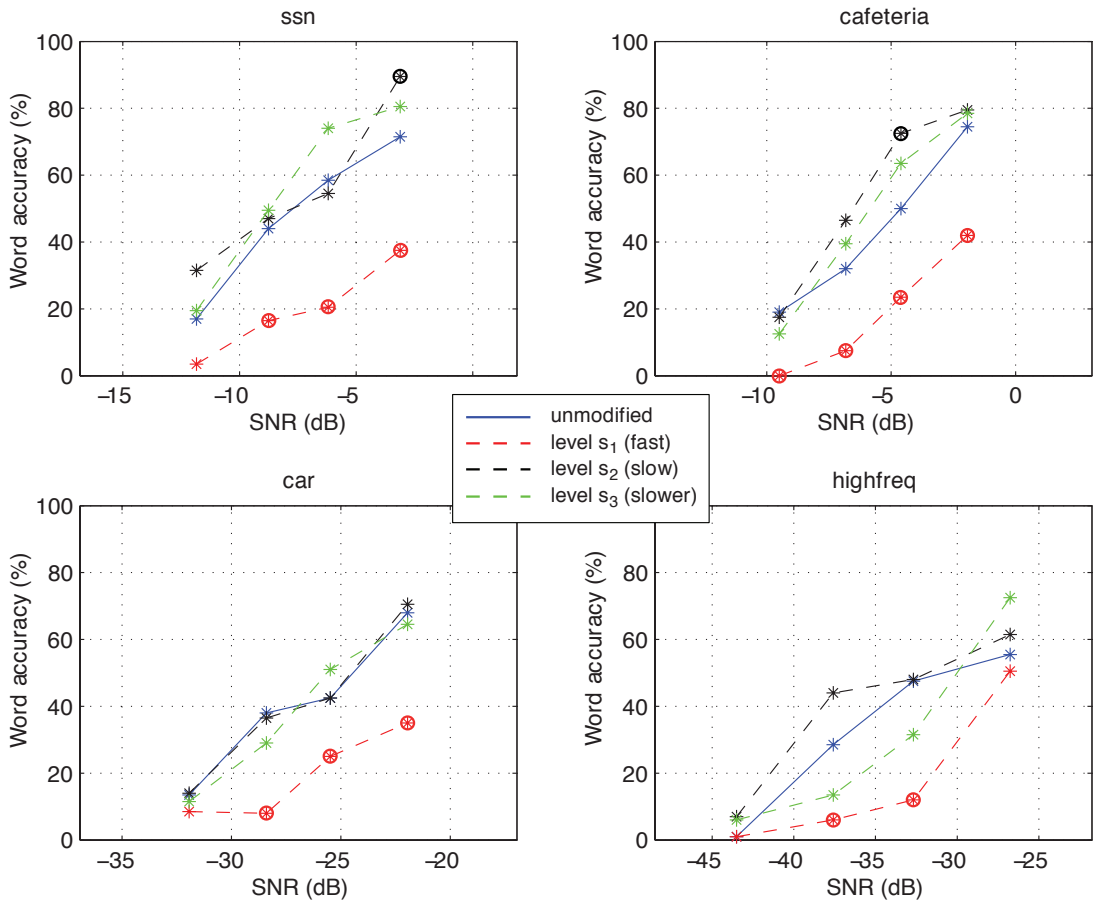


Figure 4.13: Experiment II: average subjective word accuracy for each noise type and SNR level. The curves represent different modification strengths for *speaking rate* changes. The scores significantly different to the score obtained with unmodified speech are marked with circles.

highest SNR. The modifications that had a significant impact on word accuracy were the LSP shift and speaking rate.

We obtained the largest improvements in word accuracy for the LSP shift modification, as shown in Fig. 4.12. In the presence of car noise, for the lowest SNR case, there was an improvement from 13 % to 61 % and for higher SNRs in that same noise category the word accuracy improved from 38 % to 72 % and from 42 % to 80 %. For the highest SNR level there was no significant improvement. For the ssn case there was also significant improvements from 40 % to 63 %, 59 % to 76 % and 72 % to 90 %. In the presence of cafeteria noise, the improvements were from 32 % to 61 % and 50 % to 76 %. Shifting the LSPs does not always increase intelligibility though. For high frequency noise, as we can see in Fig. 4.12, the level s_3 (i.e., the largest shift in the

LSPs) results in a significant drop in word accuracy, while the smaller shifts s_2 give a significant improvement for the higher SNR cases. This happens because the high frequency noise contains some energy in the middle frequency band, as seen in Fig. 4.3. This means that shifting LSPs by small amounts towards the high frequency region, which has the effect illustrated in Fig. 4.9, can bring intelligibility gains by boosting these mid-range frequencies.

Slower speaking rates produced significant improvements in the presence of cafeteria noise (from 50 % to 73 %), and speech-shaped noise (from 72 % to 90 %), as seen in Fig. 4.13. Unsurprisingly, speaking faster, strength s_1, always reduces intelligibility.

4.5.5 Evaluation results

We compared the performance of each measure by calculating the normalized correlation coefficient and the standard deviation of the error using the subjective score for each listening condition, averaged across listeners and sentences. In all figures and tables in this section, the measures are ordered from left to right in decreasing order of the correlation coefficient obtained for unmodified speech.

The results for each objective measure obtained for the unmodified speech are shown in Table 4.6, confirming that the measures based on auditory models outperform the other spectrum-based methods, as they did in the previous experiment. Differences between the results obtained in this experiment arise from the differences in chosen SNR values and SNR range. The spectrum-based measures IS, LSD, CEP and LLR had considerably smaller correlation coefficients. The DAU measure outperforms all measures, with a correlation coefficient of 0.94, followed by the GP measure with 0.83 and STOI with 0.79.

Fig. 4.14 shows the correlation coefficients obtained for each modification type as well as their 5 % confidence intervals. Most measures show a substantial loss in performance when speaking rate is altered, this particularly applies to the DAU, GP and STOI measures. The Dau measure obtained 0.47 for speaking rate and an average of 0.89 for the other three modifications. Smaller drops occur for the GP measure, from 0.79 to 0.52, and the STOI measure from 0.71 to 0.46. Scatter plots for the Dau and GP measures are shown in Fig. 4.15.

Table 4.7 shows the results across all modifications. There is an overall drop in the predictive power of all measures and especially for the FWS and IS measures. The GP measure had the highest correlation coefficient of 0.72, followed by the Dau mea-

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
r	0.94	0.83	0.80	0.74	0.63	0.55	0.55	0.49	0.32	0.32	0.27
σ_e	0.07	0.12	0.13	0.15	0.17	0.18	0.18	0.19	0.21	0.21	0.21

Table 4.6: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech.

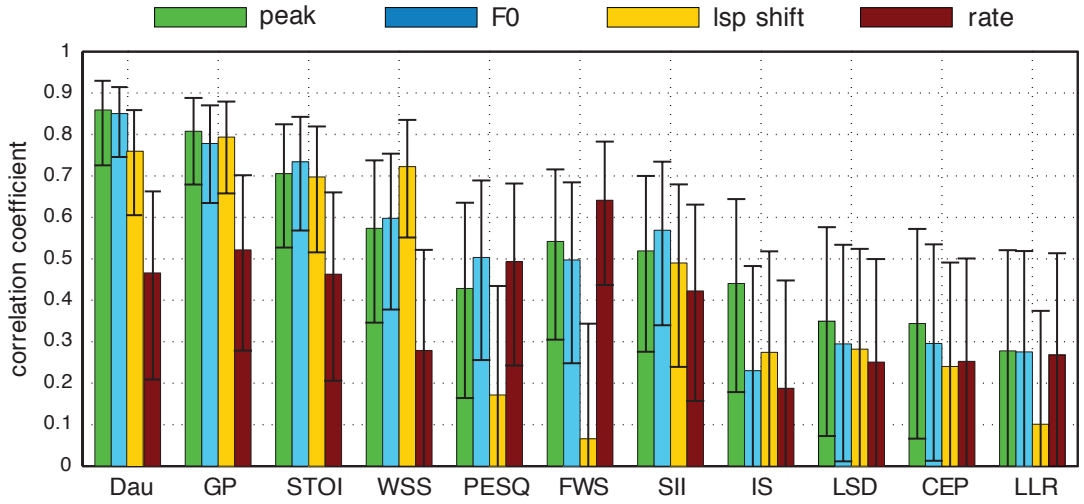


Figure 4.14: Experiment II: correlation coefficients and confidence intervals broken down by modification type. The measures are ordered in decreasing order of correlation coefficient for unmodified synthetic speech.

sure with 0.71 and STOI with 0.61. Table 4.7 also shows the correlation coefficients obtained when the unmodified speech is used as the reference signal (Case 2). The differences between the results obtained in Case 1 and Case 2 are presented for each modification type in Fig. 4.16. We can see that using the unmodified speech signal as the reference signal improves the correlation coefficient of the conventional measures IS, LSD, CEP and LLR as well as the WSS, particularly for the LSP shift modification. The Dau and the PESQ measures seem to correlate better with subjective scores when the reference signal is the modified speech and we can see that this drop is mostly due to the changes in F_0. Note that for the speaking rate modification, the modified speech was used as the reference throughout; this is because the objective measures require the reference and test signals to have the same duration.

Table 4.8 shows the correlation coefficients across unmodified speech and when the most effective modification is applied, the LSP shift. The unmodified speech signal is used as the reference speech signal because for this modification most measures have

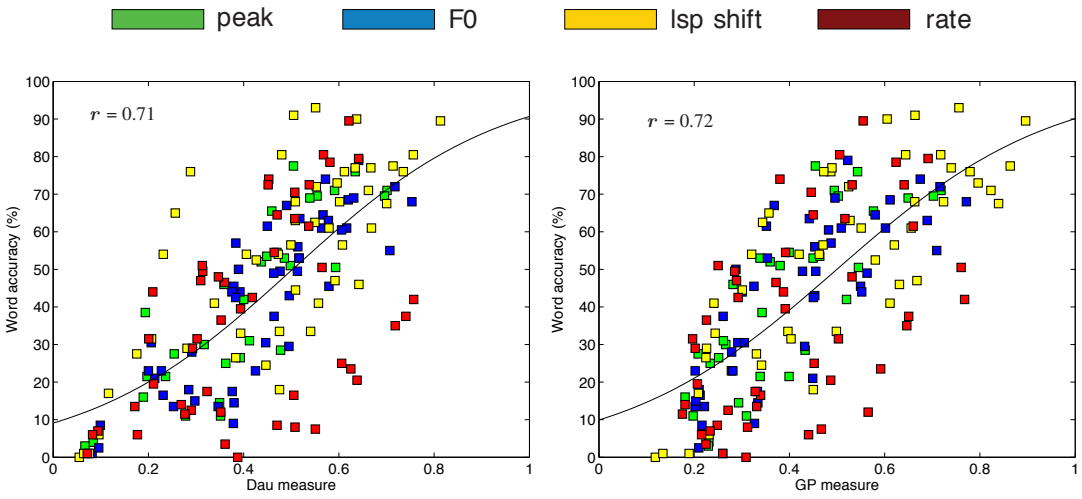


Figure 4.15: Experiment II: scatter plots of word accuracy (%) against the Dau measure (left) and GP measure (right) predictions for all modifications. Each point represents scores averaged across sentences and listeners for a certain noise condition (noise type and SNR).

better results using this reference. We can see that in this scenario the Dau measure obtained correlation of 0.77 and the GP 0.81, followed by the 0.76 obtained with the WSS. Fig. 4.17 shows the scatter plots of subjective intelligibility scores and objective scores obtained by the measures that performed the best for the more effective modification - LSP shift. Each dot represents a different listening condition combination: modification level (four in total, one for unmodified and 3 for the different LSP shift strengths), noise type (four) and noise level (four).

4.6 Discussion

The subjective intelligibility scores obtained in the two experiments with modified speech can tell us which modifications are most effective for the speech intelligibility enhancement task. In the first experiment, we applied an ideal binary mask to reallocate speech energy from time-frequency regions where it would be masked by noise to regions where it would not. This extreme reallocation strategy does not seem to be a good one as it generally did not give significant subjective intelligibility improvements across noise types and SNRs.

In our second experiment, we were interested to see whether a different class of modifications to the speech signal – Lombard-inspired modifications – have a more

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
Case 1											
r	0.71	0.72	0.61	0.48	0.35	0.13	0.45	0.16	0.27	0.25	0.17
σ_e	0.17	0.17	0.20	0.22	0.23	0.24	0.22	0.24	0.24	0.24	0.24
Case 2											
r	0.71	-	0.62	0.54	0.31	-	-	0.21	0.34	0.32	0.30
σ_e	0.17	-	0.19	0.21	0.23	-	-	0.24	0.23	0.23	0.24

Table 4.7: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *modified* synthetic speech, when we using modified speech (Case 1) or unmodified speech (Case 2) as the reference clean speech signal for calculating the objective measures that require a reference signal. The results of the measures that do not require a reference speech signal - GP, SII and FWS - are presented as belonging to Case 1.

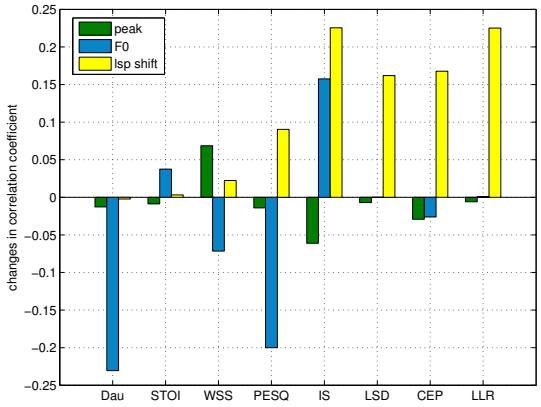


Figure 4.16: Experiment II: changes in correlation coefficients broken down by modification type when using the unmodified speech as reference.

positive effect on intelligibility and, again, which objective measures can usefully predict what the effect will be. We took generated synthetic speech and modified it at the speech parameter level to enhance spectral peaks, and to change fundamental frequency, spectral tilt and speaking rate. The subjective scores obtained indicated that the modification that increased intelligibility the most was the one that altered spectral tilt, i.e. the shift of LSPs towards higher frequencies. This modification has the effect of not only moving the formants but also of flattening the spectral tilt. We observed, however, that this modification does not *always* increase intelligibility and that the effect on intelligibility depends on the noise type and the SNR. This observation suggests that there is some optimal value of modification strength, which depends not only on

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
r	0.77	0.81	0.68	0.76	0.27	0.001	0.46	0.46	0.42	0.39	0.33
σ_e	0.16	0.15	0.18	0.16	0.24	0.25	0.22	0.22	0.23	0.23	0.24

Table 4.8: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech and *LSP shift modification*.

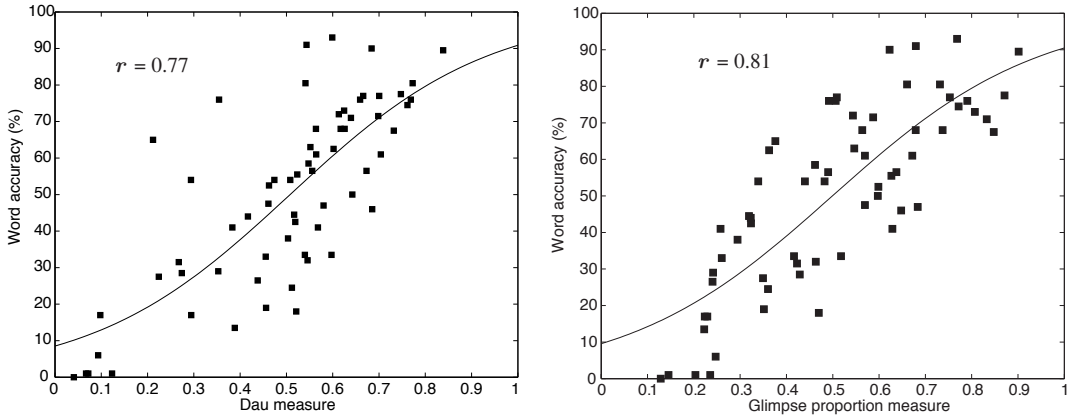


Figure 4.17: Experiment II: scatter plots of word accuracy rate (%) against the Dau measure (left) and GP measure (right) predictions for the *LSP shift modification* including the *unmodified* case. Each point represents scores averaged across sentences and listeners for a certain noise condition (noise type and SNR).

the noise type, i.e. its spectral and temporal characteristics, but also on the SNR, i.e. on the noise energy level. Although found in natural Lombard speech, spectral peak enhancement and increased F_0 did not seem to provide any significant improvement in intelligibility in the tested conditions. This last result is consistent with another study in which natural speech F_0 was modified (Lu and Cooke, 2009b). Production studies have also noted that F_0 changes are related to the increase in vocal effort: changes in F_0 are a passive result of changes of subglottal (lung) pressure and tension of the vocal folds that are required for the increase in vocal intensity (Gramming et al., 1988; Alku et al., 2002). The passive effect on F_0 of vocal effort increases (e.g., in the case of noise-induced changes) could explain why modifying F_0 does not directly impact intelligibility as observed in our experiments. Slowing the speaking rate seemed to be a good strategy only for a few combinations of noise type and level. A different modification method for changing duration, F_0 and spectral peak might however obtain positive results.

Whilst the findings regarding which modifications are most effective are interesting in themselves, they are of secondary importance in the current context. The main goal of this part of the work is to discover whether objective measures can make useful predictions about the change in intelligibility that will be brought about by various modifications to speech.

The results of the two experiments indicate that not all objective measures are suitable for the task of predicting speech intelligibility in the case where the speech is synthetic and possibly modified. In particular, the spectrum-based measure showed relatively poor performance when the noise contains energy in the higher frequency regions and when synthetic speech is modified. This means that those measures are not guaranteed to work in diverse listening conditions and would not be useful for automatically controlling the type and strength of modifications to synthetic speech. The measures that seem to perform the best across diverse listening conditions are the Dau and the GP measures; these exhibited correlation coefficients of 0.77 and 0.81 for the condition that involved unmodified speech and for the most effective modification type respectively. The predictive power of these measures is much more limited for speaking rate modifications. The explanation is most probably that changes in duration affect higher levels of processing involved in the listener's perception of speech; these cognitive levels are not taken into account by measures that attempt to model only the peripheral auditory system.

We can also remark on what the best approach is for choosing the reference signal to be used by those objective measures that require one (refer back to Fig. 4.2, results in Table 4.4 and Fig. 4.16). When a modification does not have a significant impact on intelligibility, as it is the case for fundamental frequency changes and spectral peak enhancement, choosing the modified speech seems to be the best choice for measures like Dau and PESQ. For the modifications that did have a great impact on intelligibility scores, positive or not, (IBM and LSP shift) choosing the unmodified signal as the reference signal can bring significant improvements in prediction, particularly for the measures WSS, IS, CEP, LSD and LLR. One might think that measures that use a reference signal (like the Dau measure) would perform better than ones based only on the audibility of speech in noise (like GP) as they can predict the impact of both modification and additive noise. The GP measure however obtained either better or similar results on predicting the intelligibility of modified speech in noise for all modifications we tested, see Table 4.4 and Fig. 4.16.

Improving the measures that we evaluated is out of the scope of this work, but the

cases where the measures failed can tell us something about their limitations, particularly for changes in speaking rate. For the measures that do require a reference speech signal, we believe that results for speaking rate changes could have been better if unmodified speech had been used as a reference, but this would require a time alignment because all measures assume that the processed and reference signals are of the same length. That is, none of these measures can predict the effect of changes in duration on the intelligibility of clean or noisy speech. We also believe that measures that do not require a reference speech signal would perform better on speaking rate changes if they would carry out a summation over time rather than an average of the audibility level calculated for each time frame. This should account for the increase or decrease in intelligibility that is observed when speaking rate is made slower or faster.

4.7 Conclusion

We have presented two experiments designed to evaluate the predictive power of objective measures on the intelligibility of HMM-generated synthetic speech in additive noise listening conditions. A wide variety of objective measures from several categories were used, ranging from conventional spectrum-based measures to recently-proposed measures based on rather complex models of the human auditory system. We described how we wish to use these measures to automatically control the type and degree of modification in a speech intelligibility enhancement system.

The main findings of this work are that model-based measures – notably Dau and GP – have the highest predictive power under diverse listening conditions of varying noise type and speech modification type. We also found that simple modifications at a spectral level – notably shifting LSPs – can have a significant positive impact on the intelligibility of HMM-generated synthetic speech in noise. By combining a modification strategy that improves intelligibility with an objective measure that accurately predicts the effect of that modification, we will arrived at a first version of what we were aiming for: automatically-controlled speech intelligibility enhancement.

Chapter 5

Cepstral extraction using the glimpse proportion measure

We saw in the last chapter how effective it is to modify the spectral envelope of speech to enhance TTS intelligibility in noise. Additionally, we saw that the Glimpse proportion measure highly correlates with subjective intelligibility scores of modified and unmodified TTS. In this chapter, we introduce a new cepstral extraction method based on an intelligibility measure for speech in noise, the Glimpse proportion measure. We first explain in more details how this measure operates then show how the measure can be integrated into an existing spectral envelope parameter extraction method commonly used in the HMM-based speech synthesis framework. We present how this new method changes the modeled spectrum according to the characteristics of the noise and show results for vocoded and HMM synthetic speech. Part of this chapter was published in Valentini-Botinhao et al. (2012a).

5.1 Introduction

We showed in the previous chapter that simple changes in the spectral domain can result in significant gains in intelligibility for HMM-generated synthetic speech in noise. In the same study, we also evaluated which intelligibility measures can predict these intelligibility gains. Looking at the results for unmodified and modified synthetic speech the measures that performed best were: Dau and GP. The calculation of the Dau measure involves an additional processing step – the auditory nerve response see Section 4.2.4. The GP and Dau, however, achieved comparable results. Taking into consideration both performance and simplicity, we select the Glimpse Proportion (GP) measure

(Cooke, 2006) as the most appropriate measure for the task of speech intelligibility enhancement. Our idea in this chapter is to modify the spectral envelope of speech by using the Glimpse proportion measure. To do this, we alter the optimization criterion of a cepstral coefficient extraction method commonly used in the HMM-based synthesis framework (Tokuda et al., 1995). The optimization criterion of the new cepstral extraction method proposed here, referred to as GP-based cepstral extraction, takes into account not only the mismatch between the higher dimensional spectral envelope and the cepstral-generated one but also the intelligibility in noise as given by the GP measure.

Sections 5.2 and 5.3 of this chapter provide the essential background knowledge needed to understand how to implement this idea: the maximum likelihood-based cepstral coefficient analysis method and the intelligibility measure. Section 5.4 shows how we propose to reformulate the Glimpse measure for use as a cost function for cepstral extraction. In Section 5.5, we define the proposed GP-based cepstral extraction method and show how to solve this new optimization problem. Section 5.6 presents experimental results on the acoustic analysis of the modifications and intelligibility evaluation of vocoded and HMM-generated synthetic speech. We then draw conclusions based on the obtained experimental results and point to the next steps to be taken.

5.2 Maximum likelihood-based cepstral analysis

In this section, we explain the cepstral extraction method proposed by Tokuda et al. (1995). First, we show how this method relates to the idea behind the Unbiased Estimator of the Log Spectrum (UELS) (Imai and Furuichi, 1988). We then derive the actual cost function this method minimizes and show how to solve the optimization problem.

The cepstral coefficient extraction method described in Tokuda et al. (1995) was first proposed for the extraction of cepstral coefficients and further extended to other spectral parameters like Mel cepstral (Fukada et al., 1992), generalized cepstral (Tokuda et al., 1989) and Mel generalized cepstral coefficients (Tokuda et al., 1994). It is commonly used to extract spectral parameters for training the models of an HMM-based speech synthesizer.

Tokuda et al. (1995) also proposed an adaptive version of the method using an instantaneous estimate for the gradient. Here we only show the Steepest Descent and the Newton Raphson solutions that use the real value of the gradient and therefore

achieve better results than using the instantaneous estimate.

5.2.1 Unbiased estimator of the log spectrum

The method proposed by Imai and Furuichi (1988) gives a solution to the unbiased estimation problem of the log spectrum given the speech periodogram. In this subsection, we show how the authors reached the proposed cost function for the estimator.

The authors define the modified periodogram of a wide-sense stationary process $s(n)$ as:

$$I_N(\omega) = \frac{|\sum_{n=0}^{N-1} w(n)s(n)e^{-j\omega n}|^2}{\sum_{n=0}^{N-1} w^2(n)} \tag{5.1}$$

where $w(n)$ is a window function and $s(n)$ is the speech signal waveform. Under the assumption that the modified periodogram is asymptotically unbiased we can represent it as:

$$I_N(\omega) = (1+\xi(\omega))|H(e^{j\omega})|^2 \quad \text{for } N \to \infty \tag{5.2}$$

$$E[\xi(\omega)] = 0 \tag{5.3}$$

where $|H(e^{j\omega})|$ is the magnitude spectrum of $s(n)$ and $\xi(\omega)$ is a stochastic function. We can reformulate this as:

$$E[\xi(\omega)] = E\left[\frac{I_N(\omega)}{|H(e^{j\omega})|^2} - 1\right] \tag{5.4}$$

$$= E[\exp R(\omega) - 1] \tag{5.5}$$

where

$$R(\omega) = \log I_N(\omega) - \log|H(e^{j\omega})|^2 \tag{5.6}$$

The stochastic function $\xi(\omega)$ has a uniform distribution over the frequency ω. This allows us to replace the stochastic expectation operation by frequency domain averaging:

$$E[\xi(\omega)] = E[\exp R(\omega) - 1] \tag{5.7}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}\{\exp R(\omega) - 1\}\, d\omega \tag{5.8}$$

The authors define the residue of the unbiased error as the evaluation criterion for the

estimator. The residue of the unbiased error is:

$$\rho(\omega) = \int_0^{R(\omega)} \xi(\omega)\, dX \tag{5.9}$$

$$= \int_0^{R(\omega)} \exp(X) - 1\, dX \tag{5.10}$$

$$= \exp R(\omega) - R(\omega) - 1 \tag{5.11}$$

Replacing once again the stochastic expectation operation by frequency domain averaging, the evaluation criterion is given by:

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \rho(\omega)\, d\omega \tag{5.12}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\exp R(\omega) - R(\omega) - 1\}\, d\omega \tag{5.13}$$

This evaluation criterion is the same as the Itakura-Saito distance as seen in the last chapter, or the Itakura-Saito error evaluation for AR-model in the maximum likelihood (ML) estimation method (Gray and Markel, 1976).

5.2.2 Cepstral coefficient extraction using UELS

Here we show some of the derivations found in the original paper that proposed this extraction method (Tokuda et al., 1995).

The cepstral coefficients $\{c_m\}_{m=0}^{M}$ define the spectrum of the speech signal $s(n)$ in the following way:

$$H(e^{j\omega}) = \exp \sum_{m=0}^{M} c_m e^{-jm\omega} \tag{5.14}$$

$$= K \exp \sum_{m=1}^{M} c_m e^{-jm\omega} \tag{5.15}$$

$$= KD(e^{j\omega}) \tag{5.16}$$

where $K = \exp c_0$ and $D(e^{j\omega})$ is the gain normalized version of $H(e^{j\omega})$.

We can obtain the cepstral coefficients by minimizing the criterion defined earlier in Eq.(5.13) for the unbiased condition:

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\exp R(\omega) - R(\omega) - 1\}\, d\omega \tag{5.17}$$

Since $H(e^{j\omega})$ as defined in Eq.(5.14) is a minimum phase system it is possible to prove that minimizing E with respect to $\{c_m\}_{m=1}^{M}$ is the same as minimizing the following

cost function:

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\, d\omega \tag{5.18}$$

Likewise by setting the derivative $\partial E/\partial c_0$ to zero we find that at this point:

$$K = \sqrt{\varepsilon_{min}} \tag{5.19}$$

5.2.2.1 Solving the optimization problem

Here we show how to update the vector of cepstral coefficients $\boldsymbol{c} = [c_1\ c_2\ \ldots\ c_M]^\top$ when using the Steepest Descent and Newton Raphson methods. The two methods approximate the cost function by means of a Taylor series expansion of first order (Steepest Descent) and second order (Newton Raphson).

The update formula for the cepstral coefficients vector is:

$$\boldsymbol{c}^{(i+1)} = \boldsymbol{c}^{(i)} + \mu\boldsymbol{\Delta}c^{(i)} \tag{5.20}$$

where μ is the stepsize, $\boldsymbol{\Delta}c^{(i)}$ the increment vector and i is the index for the iteration. The calculation of the increment vector for both Steepest Descent and Newton Raphson involves the gradient vector, so let us first show how we can calculate the gradient vector:

$$\boldsymbol{\nabla}\varepsilon = \frac{\partial\varepsilon}{\partial\boldsymbol{c}} \tag{5.21}$$

$$= -2\boldsymbol{r} \tag{5.22}$$

$$= -2[r_1\ r_2\ \ldots\ r_M]^\top \tag{5.23}$$

where

$$r_m = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_N(\omega)}{|D(e^{j\omega})|^2}\, e^{jm\omega} d\omega\,, \qquad m = 1,\ldots M \tag{5.24}$$

The increment vector of the Steepest Descent method is:

$$\boldsymbol{\Delta}c^{(i)} = -\boldsymbol{\nabla}\varepsilon\Big|_{\boldsymbol{c}=\boldsymbol{c}^{(i)}} \tag{5.25}$$

The increment vector of the Newton-Raphson method is:

$$\boldsymbol{\Delta}c^{(i)} = -\mathbf{H}^{-1}\boldsymbol{\nabla}\varepsilon\Big|_{c=c^{(i)}} \tag{5.26}$$

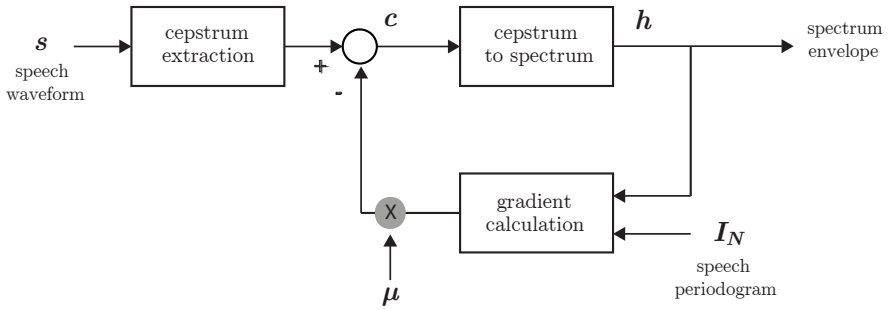


Figure 5.1: ML-based cepstral coefficient extraction using Steepest Descent.

The Hessian matrix elements are $\{\mathbf{H}\}_{i,j} = 2r_{i-j} + 2r_{i+j}$. In this report we will focus on the Steepest Descent solution as shown in Fig. 5.1. From now on we adapt the notation $H(\omega)$ to represent the discrete spectrum previously represented in the continuous domain as $H(e^{j\omega})$. The discrete frequency magnitude spectrum is given by:

$$|H(\omega_k)| = \exp \sum_{m=0}^{M} c_m \cos(m\,\omega_k) \tag{5.27}$$

where $k = 1 \ldots N$ is the index that covers the linear frequency scale uniformly. We will also adopt the following vector notation as seen in Fig. 5.1:

$\boldsymbol{s}$ = $\left[s(n)\ s(n-1)\ \ldots\ s(n-N)\right]^\top$
vector Nx1 - speech signal waveform

$\boldsymbol{h}$ = $\left[|H(\omega_1)|\ \ldots\ |H(\omega_N)|\right]^\top$
vector Nx1 - magnitude spectrum of windowed speech signal $\boldsymbol{s}$

$\boldsymbol{I_N}$ = $\left[I_N(\omega_1)\ \ldots\ I_N(\omega_N)\right]^\top$
vector Nx1 - periodogram of windowed speech signal $\boldsymbol{s}$

N scalar - number of samples in the analysis window

$\boldsymbol{c}$ = $\left[c_1\ c_2\ \ldots\ c_M\right]^\top$
vector Mx1 - cepstral coefficients

5.3 The glimpse proportion measure

The Glimpse Proportion measure was originally proposed in the context of the Glimpse model for speech perception in noise (Cooke, 2003). The model was motivated by the ability of humans to obtain information from those time-frequency regions where speech is less masked by noise and therefore less distorted (Cooke, 2003).

The GP measure (Cooke, 2006) is based on this concept: in a noisy environment, humans focus their auditory attention on 'glimpses' of speech that are not masked

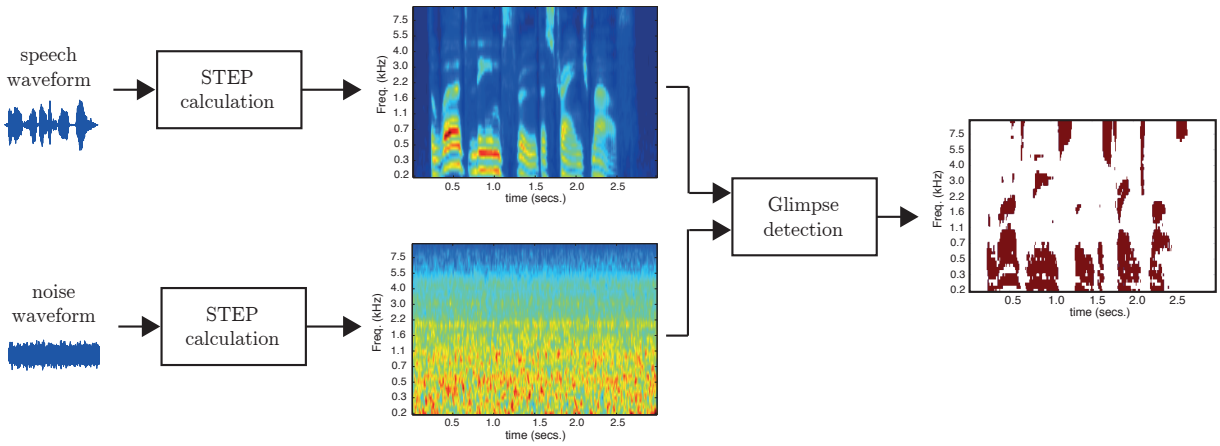


Figure 5.2: The glimpse proportion measure calculation. The measure is the percentage of glimpses detected over the entire time-frequency representation of the signals.

by noise. Rather than being a correlation, a distance or a ratio, the GP is based on audibility of speech in noise. To measure the number of available glimpses of a given speech signal in a given noise, we need the speech and noise signals to be available separately.

The GP correlates well with subjective scores for intelligibility of natural speech in noise (Cooke, 2006). In the experiments presented in the previous chapter, we also observed similar behaviour for the intelligibility of HMM-generated speech in noise even when that speech has been modified. In that experiment, we modified parameters such as the fundamental frequency (F_0) and spectral tilt to emulate Lombard speech properties; even under such modifications, GP was a reasonable intelligibility predictor (correlation coefficient above 0.8 for the most effective modification). In all these different scenarios, GP outperformed most other measures in terms of accurate predictions of intelligibility of speech in noise. An attractive property of GP is that its implementation can potentially be performed on a frame-by-frame basis as opposed to the STOI and the Dau.

The GP measure is simply the proportion of spectro-temporal regions, so called 'glimpses', where speech is more energetic than noise. To detect such glimpse the spectro-temporal excitation pattern (STEP) representation of speech and noise is compared, as shown in Fig. 5.2. To represent a signal in terms of STEP – see Fig. 5.3 – we first decompose its waveform into different frequency channels using a Gammatone filterbank whose central frequencies are linearly spaced on the Equivalent Rectangular Bandwidth (ERB) scale (Moore and Glasberg, 1996). For each channel, the temporal envelope is extracted with an absolute value operation, smoothed with a low pass

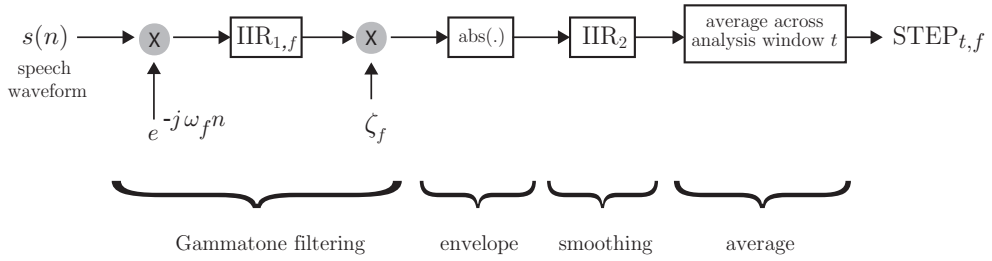


Figure 5.3: The spectro temporal excitation pattern (STEP) calculation used in the GP measure.

filter and then averaged across limited time intervals. A glimpse is detected in a time frequency region when the speech STEP value in that region is higher than the noise value.

The parameters that define the GP measure and their values are: the range of the Gammatone filters' centre frequencies (100-7500 Hz), the number of Gammatone filters N_f (55 filters), the temporal integration time for the smoothing filter τ (8 ms), the size of the time frame (30 ms) and its period (10 ms).

The centre frequencies of the Gammatone filters are linearly spaced on the ERB scale and are defined as $\tilde{w}_f$ on the ERB scale and ω_f in Hz.

The frequency response $\mathrm{IIR}_{1,f}$ of the Gammatone filter for frequency channel f is given by:

$$\mathrm{IIR}_{1,f}(z) = \frac{1 + 4a_f z^{-1} + 4a_f^2 z^{-2}}{1 - 4a_f z^{-1} + 6a_f^2 z^{-2} - 4a_f^3 z^{-3} + a_f^4 z^{-4}} \tag{5.28}$$

where $a_f = e^{-1.019\tilde{w}_f \frac{2\pi}{F_s}}$.

Cooke (1993) designed this digital filter using the impulse invariant transform method. This means that this filter is the fourth order approximation of the $\mathcal{Z}$-transform of the sampled version of the gammatone actual impulse response.

The gain ζ as shown in Fig. 5.3 normalizes the filter response gain across filters, defined as:

$$\zeta_f = \frac{[1.019\ \tilde{w}_f\ \frac{2\pi}{F_s}]^4}{3} \tag{5.29}$$

The smoothing filter is defined by the τ value as:

$$\mathrm{IIR}_2(z) = \frac{1 - e^{-\frac{1}{\tau F_s}}}{1 - e^{-\frac{1}{\tau F_s}} z^{-1}} \tag{5.30}$$

5.4 Proposed GP approximation

In this section, we show how we can approximate the GP measure so that it is completely defined by the short term magnitude spectrum of speech and consequently by the sequence of cepstral coefficients. To obtain a closed and differentiable formula that relates spectral parameters to the GP measure we make the following approximations and correspondences:

- the input signals are no longer the signal waveforms of speech and noise but the short term magnitude spectrum calculated from the short-time cepstral coefficients of speech and from the short-time discrete Fourier transform of noise (approximation)
- the previous approximation implies that all operations are carried out in the frequency domain rather than the time domain (correspondence)
- the filtering operations in the time domain are replaced by multiplications in the frequency domain with a truncated version of the frequency responses of the infinite impulse response filters (approximation due to the truncation)
- the absolute value in the time domain is replaced by a power operation that can be represented in the frequency domain as the circular convolution operation (approximation)
- the hard threshold detection of glimpses is replaced by a soft decision threshold defined by a sigmoid function (generalization).

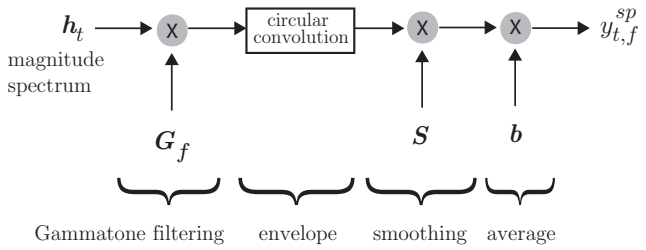


Figure 5.4: Proposed approximation for STEP calculation.

The proposed approximated GP measure is then given by:

$$GP = \frac{100}{N_f N_t} \sum_{t=1}^{N_t} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{5.31}$$

where the following scalars are defined as:

$y_{t,f}^{sp}$ STEP approximation for speech at analysis window t and frequency channel f

$y_{t,f}^{ns}$ STEP approximation for noise at analysis window t and frequency channel f

N_t number of time frames

N_f number of frequency channels

t analysis window index

f frequency channel index

$\mathcal{L}(.)$ a logistic sigmoid function defined as:

$$\mathcal{L}(x) = \frac{1}{1+e^{-\eta x}} \tag{5.32}$$

where η defines the slope of the curve. The STEP approximation as seen in Fig. 5.4 is given by:

$$y_{t,f}^{sp} = \frac{1}{N}(\boldsymbol{G}_f \boldsymbol{h}_t \circledN \boldsymbol{G}_f \boldsymbol{h}_t)^\top \boldsymbol{S}\boldsymbol{b} \tag{5.33}$$

where:

N number of frequency bins of the spectrum

$\boldsymbol{h}_t = \left[|H_t(\omega_1)| \ \ldots \ |H_t(\omega_N)| \right]^\top$
vector Nx1 - magnitude spectrum of windowed speech signal $\boldsymbol{s}$ at analysis window t

$\boldsymbol{G}_f = \text{diag}\left(\left[g_{f,1} \ \ldots \ g_{f,N} \right] \right)$
matrix NxN - diagonal matrix, diagonal contains the Gammatone filter frequency response for frequency channel f

$\boldsymbol{S} = \text{diag}\left(\left[\mathsf{s}_1 \ \ldots \ \mathsf{s}_N \right] \right)$
matrix NxN - diagonal matrix, diagonal contain the frequency response of the smoothing filter

$\boldsymbol{b} = \left[b_1 \ \ldots \ b_N \right]$
vector Nx1 - coefficients of average filter

$\circledN$ circular convolution operation dimension N

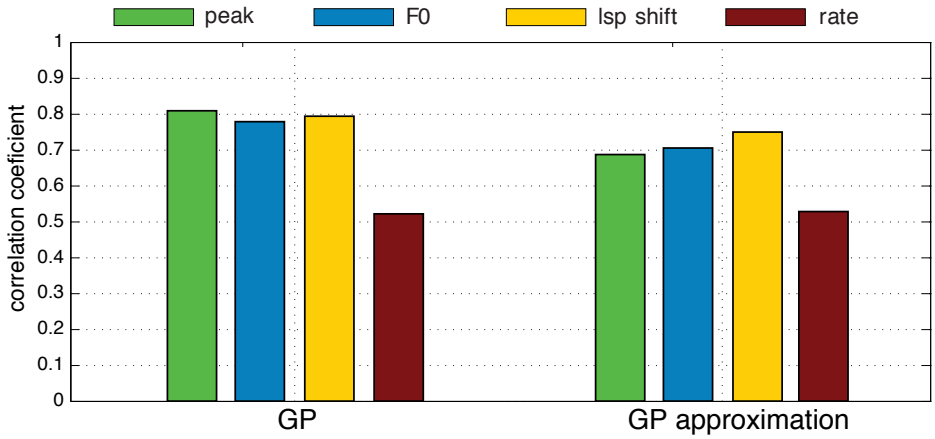


Figure 5.5: Correlation coefficients broken down by modification type obtained by the original GP measure and the GP approximation on the modified synthetic speech data described in the previous chapter in Section 4.5.

5.4.1 Evaluation of the proposed GP measure

The approximation for the GP measure that we just proposed turns the GP into a measure that depends only on the spectral envelope, in this case modelled by cepstral coefficients. As we saw in the previous chapter, spectrum-based measures poorly correlate with subjective intelligibility scores. Here we calculate the correlation coefficient of the GP approximation in Experiment II for each modification. These results are shown in Fig. 5.5. We can see that the GP approximation is not as strongly correlated to subjective scores as the original measure for the peak enhancement and fundamental frequency modifications. The drop in correlation is much smaller for the LSP shift modification and we found that the GP approximation is as good a predictor for speaking rate changes as the original GP. Compared to the other measures (see Fig. 4.14) the GP approximation is still a much better predictor of modified synthetic speech, even though it is now a spectrum-based measure.

5.5 GP-based cepstral coefficient extraction

In this section, we show how to integrate the approximated GP measure shown in the previous section into the existing cost function for cepstral coefficient extraction shown in Section 5.2.

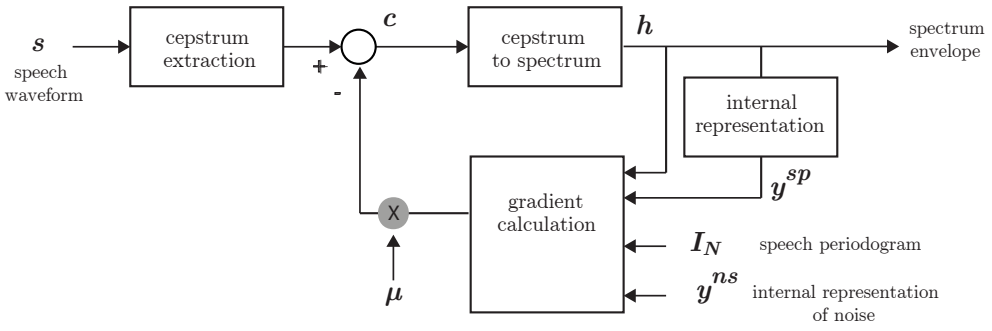


Figure 5.6: GP-based cepstral coefficient extraction using Steepest Descent.

5.5.1 Cost function

In order to keep a compromise between the minimization of the cost function defined in Eq.(5.18) and the maximization of the intelligibility measure given by Eq.(5.31) we need to define an extra parameter that controls the weight given to each criterion. This parameter is called β. We can then redefine the cost function as:

$$E_t = \varepsilon_t - \beta GP_t \tag{5.34}$$

where

$$GP_t = \frac{100}{N_f} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \tag{5.35}$$

The spectral parameter vector $\boldsymbol{c}_t = [c_{t,1} \quad c_{t,2} \ldots \quad c_{t,M}]^\top$ is then given by:

$$\boldsymbol{c}_t^* = \operatorname{argmin} E_t(\boldsymbol{c}_t) \tag{5.36}$$

$$= \operatorname{argmin} \left[\varepsilon_t(\boldsymbol{c}_t) - \beta GP_t(\boldsymbol{c}_t)\right] \tag{5.37}$$

It is clear that when β=0.0 the GP-based cepstral extraction method becomes the original cepstral coefficient extraction method of Section 6.2.

5.5.2 Steepest descent solution

Fig. 5.6 shows a block diagram representation of how we update the cepstral coefficients iteratively using the Steepest Descent method given the spectral envelope $\boldsymbol{h}$, periodogram $\boldsymbol{I}_N$ and internal representations of speech $\boldsymbol{y}^{sp}$ and noise $\boldsymbol{y}^{ns}$.

The update equation for cepstral coefficients using Steepest Descent is:

$$\boldsymbol{c}^{(i+1)} = \boldsymbol{c}^{(i)} + \mu \boldsymbol{\Delta} c^{(i)} \tag{5.38}$$

$$= \boldsymbol{c}^{(i)} - \mu \boldsymbol{\nabla} E_t^{(i)} \tag{5.39}$$

where $\mu = 1/||\nabla E_t^{(i)}||$. The gradient vector is:

$$\nabla E_t^{(i)} = \nabla \varepsilon_t^{(i)} - \beta \nabla GP_t^{(i)} \tag{5.40}$$

The gain is still updated as:

$$K^{(i)} = \sqrt{\varepsilon_{min}^{(i)}} \tag{5.41}$$

From now on we will drop the iteration index (i) for clarity. To solve the optimization using Steepest Descent we need to calculate the new term of the gradient defined in Eq.(5.40), that is the gradient of the GP function:

$$\nabla GP_t = \frac{\partial GP_t}{\partial \boldsymbol{c}_t} = \frac{100}{N_f} \sum_{f=1}^{N_f} \frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial \boldsymbol{c}_t} \tag{5.42}$$

$$= \frac{100}{N_f} \sum_{f=1}^{N_f} \frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} \tag{5.43}$$

We can write the first term in the summation of Eq.(5.43) as:

$$\frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} = \eta\, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \left[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\right] \tag{5.44}$$

The second term in the summation of Eq.(5.43) is given by:

$$\frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} = \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{h}_t} \tag{5.45}$$

The first term on the right side of Eq.(5.45) is a matrix of dimension MxN defined as:

$$\boldsymbol{H}_{c_t} \equiv \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} = \begin{bmatrix} \frac{\partial |H_t(\omega_1)|}{\partial c_{t,1}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,1}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,1}} \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,2}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,2}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,2}} \\ \vdots & & & \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,M}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,M}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,M}} \end{bmatrix}$$

When the spectrum is modeled by cepstral coefficients as in Eq.(5.27) the elements of this matrix are:

$$\{\boldsymbol{H}_{c_t}\}_{m,k} = \frac{\partial |H_t(\omega_k)|}{\partial c_{t,m}} = |H_t(\omega_k)| \cos(m\,\omega_k) \tag{5.46}$$

where k is the index for the spectrum frequency bin and m as defined previously is the index for the cepstral coefficients.

The second term of Eq.(5.45) depends on the definition of the STEP approximation in Eq.(5.33) and it is then given by:

$$\frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{h}_t} = \frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{h}_t} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{l}_{t,f}} \tag{5.47}$$

$$= \frac{1}{N} \frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{h}_t} \boldsymbol{S}\boldsymbol{b} \tag{5.48}$$

$$= \frac{1}{N} \frac{\partial \boldsymbol{G}_f \boldsymbol{h}_t}{\partial \boldsymbol{h}_t} \frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{G}_f \boldsymbol{h}_t} \boldsymbol{S}\boldsymbol{b} \tag{5.49}$$

$$= \frac{1}{N} \boldsymbol{G}_f \frac{\partial \boldsymbol{l}_{t,f}}{\partial \boldsymbol{G}_f \boldsymbol{h}_t} \boldsymbol{S}\boldsymbol{b} \tag{5.50}$$

$$= \frac{1}{N} \boldsymbol{G}_f \left(2\boldsymbol{\Gamma}_N \circledN \boldsymbol{G}_f \boldsymbol{h}_t\right) \boldsymbol{S}\boldsymbol{b} \tag{5.51}$$

where $\boldsymbol{l}_{t,f} = (\boldsymbol{G}_f \boldsymbol{h}_t \circledN \boldsymbol{G}_f \boldsymbol{h}_t)$ and $\boldsymbol{\Gamma}_N$ is the identity matrix of dimension N.

The operation $(\boldsymbol{\Gamma}_N \circledN \boldsymbol{G}_f \boldsymbol{h}_t)$ defines a matrix NxN of the following form:

$$\begin{bmatrix} \boldsymbol{e}_1 \circledN (\boldsymbol{G}_f \boldsymbol{h}_t)^\top \\ \boldsymbol{e}_2 \circledN (\boldsymbol{G}_f \boldsymbol{h}_t)^\top \\ \vdots \\ \boldsymbol{e}_N \circledN (\boldsymbol{G}_f \boldsymbol{h}_t)^\top \end{bmatrix}$$

where $\boldsymbol{e}_n$ is the n-th column of the identity matrix $\boldsymbol{\Gamma}_N$.

Connecting eqs.(5.44), (5.46) and (5.51), the gradient vector is given by:

$$\boldsymbol{\nabla} GP_t = \frac{100}{N_f N} \sum_{f=1}^{N_f} \eta \, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \left[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\right] \boldsymbol{H}_{c_t} \boldsymbol{G}_f \left(2\boldsymbol{\Gamma}_N \circledN \boldsymbol{G}_f \boldsymbol{h}_t\right) \boldsymbol{S}\boldsymbol{b} \tag{5.52}$$

Algorithm using Steepest Descent

Given the speech periodogram $\boldsymbol{I}_N$ and the noise waveform, for each time frame t:

- Calculate internal representation of noise: $\boldsymbol{y}_t^{ns} = \left[y_{t,1}^{ns} \ \dots \ y_{t,N_f}^{ns} \right]^\top$
- Initialize $[K_t \ \boldsymbol{c}_t]$ as the first $M+1$ values of the minimum-phase cepstrum: $\mathcal{F}^{-1}\left\{ 0.5 \log I_N(\omega) \right\}$, where $\mathcal{F}^{-1}$ is the inverse discrete Fourier transform operation.
- Optimization loop:
  - Given the new spectral envelope $\boldsymbol{h}_t$ calculate the speech internal representation $\boldsymbol{y}_t^{sp} = \left[y_{t,1}^{sp} \ \dots \ y_{t,N_f}^{sp} \right]^\top$
  - Calculate gradient vector ∇E_t
  - Update $\boldsymbol{c}_t$, set $K_t = \sqrt{\varepsilon_{min}}$ and calculate the new spectral envelope $\boldsymbol{h}_t$
  - If converges or distortion is above threshold then stop.

5.5.3 Energy normalization

In this section we explain how to reformulate the optimization problem in order to keep the overall energy of speech constant. For clarity reasons we drop the time index t in the equations and use the continuous representation of the spectrum $H(e^{j\omega})$.

If the 0-th cepstral coefficient is not modified we can not guarantee that the energy is kept constant because this coefficient does not represent the energy of the spectral envelope but the log-energy:

$$c_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |H(e^{j\omega})|^2 \, d\omega \tag{5.53}$$

Let us first define the quantity we refer here as overall energy of speech in a certain time frame:

$$\sum_{n=0}^{N-1} |s(n)|^2 = \psi \tag{5.54}$$

where N is the size of the time window frame. From Parseval we have that:

$$\sum_{n=0}^{N-1} |s(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \, d\omega = \psi \tag{5.55}$$

where $S(e^{j\omega})$ is the discrete time Fourier transform of time signal $s(n)$.

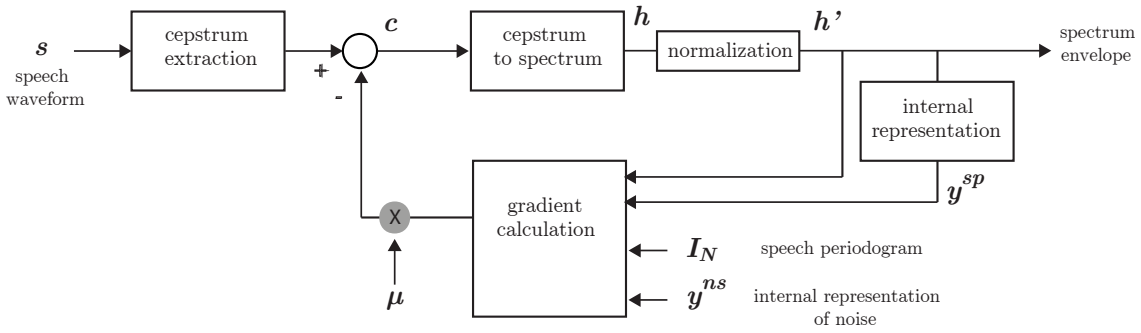


Figure 5.7: GP-based cepstral coefficient extraction using Steepest Descent with energy normalization.

This can be related to the spectral envelope $H(e^{j\omega})$ and the frequency representation $E(e^{j\omega})$ of the excitation signal:

$$\psi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})E(e^{j\omega})|^2 d\omega \tag{5.56}$$

We can assume that $|E(e^{j\omega})|$ is constant over the frequency domain for both voiced and unvoiced segments. For voiced speech segments this is true if the size of the analysis window is set to two pitch periods and for unvoiced segments this is true because at these segments the excitation signal is white noise. Under this assumption and considering that the cepstral extraction method does not modify the excitation signal we can assume that in order to keep the energy in the time domain constant it is sufficient to keep the following constant:

$$\psi = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 d\omega \tag{5.57}$$

The minimization of the cost function as given in Eq.(5.37) should then be solved subject to the above constraint. Solving a nonconvex optimization problem is however a hard task. One feasible solution is to perform, at each iteration of the Steepest Descent method, an energy normalization operation and alter the objective function and consequentially the gradient vector accordingly. Fig. 5.7 shows this solution. To explain how the gradient should be modified we first need to define the operation that normalizes the energy of the spectrum.

The following operation modifies the spectrum $|H(e^{j\omega})|$ with overall energy ψ so that the resulting spectrum $|H'(e^{j\omega})|$ has an overall energy equal to ψ':

$$|H'(e^{j\omega})| = \frac{|H(e^{j\omega})|}{\sqrt{\frac{1}{\psi'}\int_{-\pi}^{\pi}|H(e^{j\omega})|^2 d\omega}} = \frac{|H(e^{j\omega})|}{\sqrt{\frac{\psi}{\psi'}}} \tag{5.58}$$

In order to modify the gradient we need to see the impact of this operation in the cepstral coefficient domain. The normalization operation transforms a set of cepstral coefficients c_m that model the spectrum $|H(e^{j\omega})|$ with overall energy ψ, into parameters c'_m that model a spectrum $|H'(e^{j\omega})|$ with overall energy equal to ψ' in the following way:

$$|H'(e^{j\omega})| = \frac{|H(e^{j\omega})|}{\sqrt{\frac{\psi}{\psi'}}} \tag{5.59}$$

$$= \frac{\exp \sum_{m=0}^{M} c_m \cos(m\omega)}{\sqrt{\frac{\psi}{\psi'}}} \tag{5.60}$$

$$= \frac{\exp \sum_{m=0}^{M} c_m \cos(m\omega)}{\exp\left[\log \sqrt{\frac{\psi}{\psi'}}\right]} \tag{5.61}$$

$$= \exp\left[\Big(\sum_{m=0}^{M} c_m \cos(m\omega)\Big) - 0.5 \log\left(\frac{\psi}{\psi'}\right)\right] \tag{5.62}$$

$$= \exp \sum_{m=0}^{M} c'_m \cos(m\omega) \tag{5.63}$$

The energy-normalized cepstral coefficients c'_m are then given by:

$$c'_m = \begin{cases} c_0 - 0.5 \log\left(\frac{\psi}{\psi'}\right) & m = 0 \\ c_m & m \neq 0 \end{cases} \tag{5.64}$$

Only the first cepstral coefficient changes so we can write:

$$|H'(e^{j\omega})| = |K'||D(e^{j\omega})| \tag{5.65}$$

where

$$K' = \exp(c'_0) \tag{5.66}$$

$$= \exp\left[c_0 - 0.5 \log\left(\frac{\psi}{\psi'}\right)\right] \tag{5.67}$$

If ψ is equal to ψ', i.e. the energy-normalization operation has no impact on the spectrum, we can see that c'_m is equal to c_m. The only term in the gradient vector ∇GP that needs to be adjusted is the one given by Eq.(5.46). To show how this term changes we adopt the discrete representation $H(\omega_1), \ldots, H(\omega_N)$ of the spectrum. Eq.(5.57) is then approximated to:

$$\psi = \sum_{k=1}^{N} |H(\omega_k)|^2 \tag{5.68}$$

With the energy normalization operation the derivative in Eq.(5.46) becomes:

$$\frac{\partial|H'(\omega_k)|}{\partial c_m} = \frac{\partial|K'||D(\omega_k)|}{\partial c_m} \tag{5.69}$$

$$= \frac{\partial|K'|}{\partial c_m}|D(\omega_k)| + |K'|\frac{\partial|D(\omega_k)|}{\partial c_m} \tag{5.70}$$

$$= |K'|\frac{\partial c'_0}{\partial c_m}|D(\omega_k)| + |K'||D(\omega_k)|\cos(m\omega_k) \tag{5.71}$$

$$= |H'(\omega_k)|\frac{\partial c'_0}{\partial c_m} + |H'(\omega_k)|\cos(m\omega_k) \tag{5.72}$$

$$= |H'(\omega_k)|\Big(\frac{\partial c'_0}{\partial c_m} + \cos(m\omega_k)\Big) \tag{5.73}$$

The derivative term in the previous equation is given by:

$$\frac{\partial c'_0}{\partial c_m} = \frac{\partial c_0}{\partial c_m} - 0.5\,\frac{\psi'}{\psi}\,\frac{1}{\psi'}\,\frac{\partial \psi}{\partial c_m} \tag{5.74}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi}\sum_{l=1}^{N}|H(\omega_l)|^2\cos(m\omega_l) \tag{5.75}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\omega_l) \tag{5.76}$$

$$\frac{\partial c'_0}{\partial c_m} = \begin{cases} 0.0 & m = 0 \\ -\frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\omega_l) & m \neq 0 \end{cases} \tag{5.77}$$

The derivative in Eq.(5.46) then becomes:

$$\frac{\partial|H'(\omega_k)|}{\partial c_m} = \begin{cases} |H'(\omega_k)| & m = 0 \\ |H'(\omega_k)|\Big(\cos(m\omega_k) - \frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\omega_l)\Big) & m \neq 0 \end{cases} \tag{5.78}$$

Comparing this equation with Eq.(5.46) we can see that the energy normalization constraint just added a new term to the equation. Using this new gradient calculation, and normalising the speech energy at each iteration, guarantees that the energy of the speech signal is fixed during gradient descent optimization. Because the optimization is performed per analysis window, the energy of each window will not change, meaning that there is no reallocation of energy across windows and that the maximisation of the GP is bounded by the amount of energy initially available in the analysis window.

Algorithm using Steepest Descent and energy normalization

Given the speech periodogram $\boldsymbol{I}_N$ and the noise waveform, for time frame t:

- Calculate internal representation of noise: $\boldsymbol{y}_t^{ns} = \left[y_{t,1}^{ns} \ \dots \ y_{t,N_f}^{ns} \right]^\top$
- Initialize $[K_t \ \boldsymbol{c}_t]$ as the first $M+1$ values of the minimum-phase cepstrum: $\mathcal{F}^{-1}\left\{ 0.5 \log I_N(\omega) \right\}$, where $\mathcal{F}^{-1}$ is the inverse discrete Fourier transform operation.
- Calculate the overall energy ψ' from the periodogram
- Optimization loop:
  - Normalize spectrum so that overall energy is $\psi' \rightarrow \boldsymbol{h}_t'$ and $\boldsymbol{c}_t'$
  - Given the new spectral envelope $\boldsymbol{h}_t'$ calculate speech internal representation $\boldsymbol{y}_t^{sp} = \left[y_{t,1}^{sp} \ \dots \ y_{t,N_f}^{sp} \right]^\top$
  - Calculate gradient vector $\boldsymbol{\nabla} E_t$
  - Update $\boldsymbol{c}_t$ and calculate $\boldsymbol{h}_t$
  - If converges or distortion is above threshold then stop.

5.6 Evaluation

To find whether the proposed method for cepstral extraction increases the number of glimpses under the original glimpse definition, we evaluate two sets of speech data: vocoded and synthetic speech. Vocoded speech is natural speech that has been vocoded, i.e. analysed into a compact parametric representation and then reconstructed. To generate modified vocoded speech, we extract cepstral parameters using the proposed GP-based method. To test synthetic speech we compare a TTS voice trained with cepstral coefficients extracted with the original method to a voice built with cepstral coefficients extracted with the proposed GP-based method. Additionally, to find whether the increase in glimpses actually resulted in an increase in subjective intelligibility score, we perform a listening test with these two types of speech material added to speech-shaped and high frequency noise – the noises that were used to drive the GP-based cepstral extraction method.

5.6.1 Stimuli

The speech material we used to generate vocoded speech was the semantically unpredictable sentences (SUS) set from the Blizzard Challenge 2010 (King and Karaiskos, 2010). The samples were of the British male speaker named *rjs* – the same speaker used to train the synthetic voice for the evaluations of the last chapter – sampled at 20kHz. To train the synthesis models we used 1000 different sentences from the same speaker also at 20kHz. The text of the sentences used to generate vocoded speech were used as test sentences for the HMM-generated synthetic speech. To generate vocoded and synthetic speech we used as synthesis filter the log spectrum approximation filter (Tokuda et al., 1995) with simple excitation as input.

We used the same set of spectral and excitation parameters to analyse natural speech for both the generation of vocoded speech and the training of the acoustic model. Using the proposed method we extracted 52 cepstral coefficients for different β values – the value that controls the weights of cost functions in Eq.(5.34) – , including the $\beta = 0$ case for comparison. The periodogram was set to be the smoothed spectrum extracted using STRAIGHT (Kawahara et al., 1999), see Section 3.1.3.2 for how STRAIGHT smoothed spectrum is extracted from speech waveform. We initialize the algorithm with the first $M+1$ values of the minimum phase cepstrum. The step size was set to $\mu^{(i)} = 1/||\nabla E_t^{(i)}||$. We used both total error convergence and maximum distortion, defined as the UELS cost function given by Eq.(5.18), as the stopping criterion.

The acoustic model that we used for synthetic speech was a hidden semi-Markov model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values. We used one stream for the spectrum and three streams for the logF0. We used the Global Variance method (Toda and Tokuda, 2007) to compensate for the oversmoothing effect of the acoustical modeling.

For these experiments, we added vocoded and HMM-generated synthetic speech to two different types of stationary noise, speech-shaped noise (SSN) and high frequency noise (HF), the same noise types used for the experiments in the previous chapter. Each masker was added at a different SNR: 0 dB for SSN and and -20 dB for HF. We compare the intelligibility of the different voices under a fixed SNR, which is computed at a sentence level. As the modification method proposed here keeps the energy level within each analysis frame fixed, no other energy normalization had to be performed to guarantee that the energy level of the sentence was not modified.

5.6.2 GP scores

We calculated the value of the GP at a sentence level and then averaged across the test set. The results of the original/proposed cepstral extraction method using vocoded speech are: 22.6/26.8 (HF) and 16.5/23 (SSN). Results for synthetic speech are: 22.2/26.2 (HF) and 15.1/19.1 (SSN). Even though the proposed method for cepstral extraction maximizes an approximated version of the GP, the results show that the original GP measure also increased.

5.6.3 Acoustic analysis

Fig. 5.8 shows the Long Term Average Spectrum (LTAS) of vocoded speech generated using the original and the proposed method when noise is speech-shaped. In the figure we can also see the LTAS of the noise. We can see that on average the proposed method reallocates energy mostly to the frequency range between 800 Hz and 4.8 kHz, the band where the auditory human system is most sensitive. The attenuation occurs mostly in the lower frequency regions below 800 Hz. Fig. 5.9 shows the LTAS of vocoded speech generated using the original and the proposed cepstral extraction method for the high frequency masker. In this noise the energy boost occurs in a similar region, but in a much smaller strength. We can also observe some attenuation in the high frequency region, as this region is highly masked by noise.

5.6.4 Listening experiment

For the listening test we played all signals over headphones to participants in sound-isolated booths. Each individual sentence could be played only once before the participant had to type in what he or she heard. A total of eight native English speakers participated in the experiment with vocoded speech and another eight participants were assigned to the experiment with synthetic speech. Each participant heard twelve different sentences per noise type.

5.6.5 Results and discussion

Fig. 5.10 shows the word accuracy rates obtained in the listening test with vocoded (left) and synthetic speech (right). Each group mean is represented by a circle; two means are significantly different at a 0.05 level only if their intervals are disjoint.

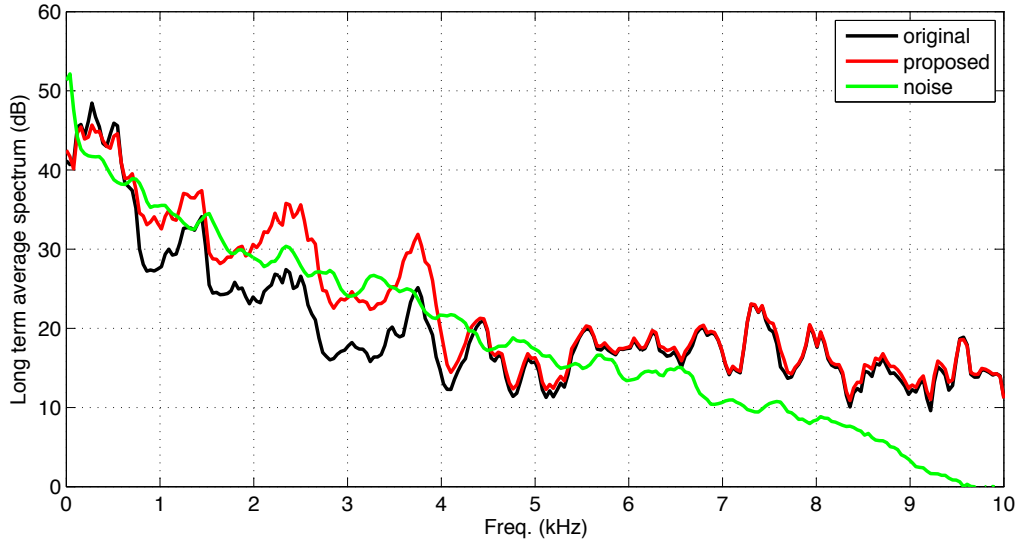


Figure 5.8: Long term average spectrum curves extracted for vocoded speech generated using the original method ($\beta = 0$) and the proposed method ($\beta \neq 0$) for speech-shaped noise at 0 dB SNR.

We can see that the proposed method does not produce any significant differences in word accuracy for vocoded speech. However for synthetic speech and speech-shaped noise there is a significant improvement of word accuracy from 31 % to 44 % (a gain of 44 % relative).

For the high frequency noise case it seems that, although not significantly, the proposed method decreases the word accuracy rates. We believe this happens because the modifications imposed by such noise leads to less natural speech which in turn degrades intelligibility.

The impact of the proposed method seems to be stronger for synthetic speech although the GP gains were smaller or similar for synthetic speech, most probably because in harder tasks smaller glimpse variations lead to stronger effects.

5.7 Conclusion

In this chapter, we showed how to use a measure of speech intelligibility in noise to modify HMM-synthetic speech and make it more intelligible for a certain noise. We proposed a new cepstral extraction method that aims not only to minimize the mismatch between periodogram and modelled spectrum but also to maximize speech intelligibility in noise, as defined by the Glimpse Proportion measure.

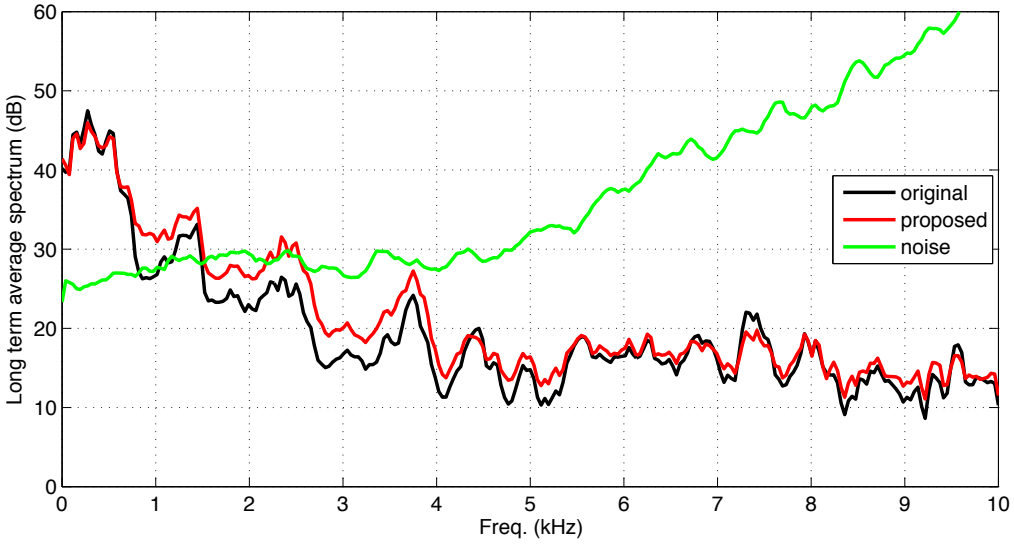


Figure 5.9: Long term average spectrum curves extracted for vocoded speech generated using the original method ($\beta = 0$) and the proposed method ($\beta \neq 0$) for high frequency noise at -20 dB SNR.

In the background sections, we explained an existing method for cepstral coefficient extraction commonly used in the HMM-based speech synthesis framework. We then explained the intelligibility measure we chose to make use of, the Glimpse Proportion measure, and why we chose this measure.

We then proposed how this measure can be reformulated to be used in the context of cepstral coefficient extraction and how to integrate it to the existing cost function. We call this new proposed method GP-based cepstral coefficient extraction. We showed how the optimization problem can be solved using the gradient information and how to keep overall energy of speech constant during optimization.

The listening tests with vocoded and synthetic speech showed the effectiveness of the method for speech-shaped noise but not for high frequency noise, which might indicate that the amount of distortion introduced into the speech by the modification was too large.

We have seen that increasing GP values does not necessarily result in intelligibility improvements. This shows us how important it is to control the strength of modification up to a certain acceptable level where the measure operates correctly. Our next step is to handle distortion in a better way as well as extending the proposed method for Mel cepstral coefficients, since the quality of HMM-synthetic speech trained with these coefficients is superior to voices trained with cepstral coefficients.

We also plan to apply a similar idea as a post processing method for modifying the

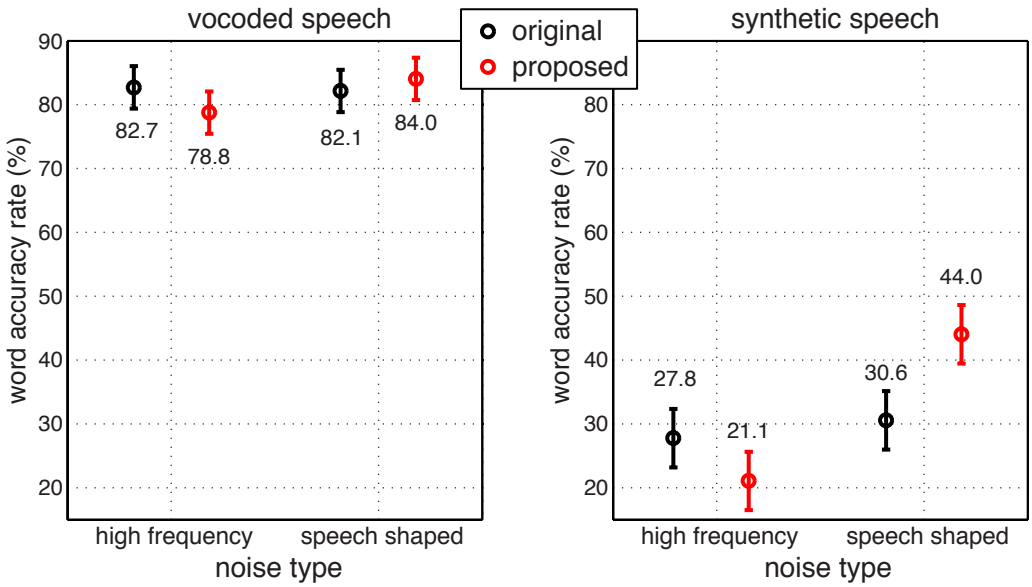


Figure 5.10: Word accuracy rates of listening test with vocoded (left) and synthetic (right) speech.

generated sequence of spectral coefficients. Under this framework, we would be able to train a single synthesis model and then further refine its inferred parameters so that intelligibility of generated speech in a given noise environment is increased. This is obviously preferable to training individual synthesis models for each noise situation. Performing intelligibility enhancement at generation time rather than at training time allows for modifications that compensate for noises that change at generation time, i.e. fluctuating noises like a competing talker situation. We will also compare the GP-based approach to other methods for intelligibility enhancement, like for instance using clear or Lombard speech data to train or adapt the models.

Chapter 6

Mel cepstral modification using the glimpse proportion measure

In Chapter 5, we proposed a method to extract cepstral coefficients that not only minimizes the errors of modelling the spectrum with a small set of cepstral coefficients but also maximizes the glimpse measure for a particular noise masker. In this chapter, we propose an alternative to this method which can be applied at generation time: a cepstral coefficient modification. This new method alters the Mel cepstral coefficients – cepstral coefficients defined on the Mel scale – in order to increase the intelligibility of the speech in the presence of a known noise. The method can operate at generation time which means that it can deal with non-stationary noises like a competing speaker. Similar to the earlier extraction method, the new method is based on the Glimpse Proportion (GP) measure approximation proposed in the previous chapter. We first show how to modify the Mel cepstral coefficients iteratively using the GP measure approximation as an optimization criterion and how to control the modification by limiting its frequency resolution. Then, to evaluate the method, we built eight different voices from normal read-text speech data from a male speaker. We present results of an acoustic analysis and subjective intelligibility scores. This work was partially published in (Valentini-Botinhao et al., 2012c,d, 2013c).

6.1 Introduction

We observed in Chapter 4 that the Glimpse Proportion (GP) measure for speech intelligibility in noise (Cooke, 2006) has a high correlation coefficient with subjective intelligibility scores for HMM-generated synthetic speech whose spectral envelope has

been modified. Moreover, modifications in the spectral envelope domain can achieve quite high intelligibility gains. In Chapter 5, we proposed a cepstral extraction method based on the GP measure for the HMM-based synthesis framework. This method was shown to provide a significant intelligibility improvement, although not for all noise types. We hypothesise that this is due to distortions introduced by the method itself. The compromise between increasing glimpses and minimizing the mismatch between spectrum and the spectral envelope as modelled by cepstral coefficients is not an easy one to attend: glimpses can be created with the introduction of audible distortions. Another disadvantage of that approach is having to train a different synthesis model for each noise type as the noise-dependent modification is performed as part of feature extraction. Now, we propose a method that can be applied at generation time, and does not require any information about the spectral envelope of natural speech to achieve distortion control. In this new method, we maximize the GP alone. The maximization of the GP without any constraint will generate glimpses across all spectral envelope generating audible distortions. To control the modification and the distortion we act in two ways: using a stopping criteria based on the mismatch between the auditory representations of modified and unmodified speech, as proposed by the GP measure, and only modifying the first few cepstral coefficients, thus limiting the frequency resolution of the modifications. A further extension proposed here is the possibility of using this method for *Mel* cepstral coefficients, which can provide higher speech quality with fewer coefficients (Imai, 1983; Fukada et al., 1992).

Although the formulation of the problem allows for the extension to other types of spectral parametrization such as the Mel Generalized Cepstral coefficients (MGC) (Koishida et al., 1996) we can not guarantee that the synthesis filter created from such modified MGCs is stable, see Section 3.1.3.2. Stability is always guaranteed for any value of Mel cepstral coefficients though. To modify the MGC parameters it would be necessary to first transform them into a representation where stability is easily ensured like the MGC-LSP as proposed in Koishida et al. (2000).

In Section 6.2, we show how Mel cepstral coefficients model the speech spectrum. In Section 6.3, we introduce the new method for Mel cepstral modification based on the GP measure. We then provide, in Section 6.4, experimental results from listening experiments to support our conclusions.

6.2 Mel cepstral coefficients

We can represent the spectrum $H(e^{j\omega})$ by a M-th order Mel cepstral coefficient set $\{c_m\}_{m=0}^{M}$ using the following equation (Fukada et al., 1992):

$$H(e^{j\omega}) = \exp \sum_{m=0}^{M} c_m e^{-jm\tilde{\omega}} \tag{6.1}$$

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{6.2}$$

where α is the warping factor that controls the frequency scaling.

We can choose α such that $\tilde{\omega}$ spans the frequency axis on a particular scale, for instance the Mel scale, creating so-called Mel cepstral coefficients (Fukada et al., 1992). When using the Mel scale warping, we can represent the spectral envelope with fewer coefficients than when using a linear frequency scale, without a loss in quality (Imai, 1983).

According to Eq.(6.1), the magnitude spectrum is defined by the Mel cepstral coefficients as follows:

$$|H(e^{j\omega})| = \exp \sum_{m=0}^{M} c_m \cos(m\tilde{\omega}) \tag{6.3}$$

From now on we adapt the notation $H(\omega)$ to represent the discrete spectrum previously represented in the continuous domain as $H(e^{j\omega})$. The previous equation becomes:

$$|H_t(\omega_k)| = \exp \sum_{m=0}^{M} c_m \cos(m\tilde{\omega}_k) \tag{6.4}$$

where $k = 1 \ldots N$ is the index that covers a frequency scale uniformly.

6.3 GP-based Mel cepstral modification

6.3.1 Cost function

Given a set of Mel cepstral coefficients and a noise signal we want to obtain a new set of Mel cepstral coefficients $\boldsymbol{c}_t = [c_{t,1} \ \ldots \ c_{t,m} \ \ldots \ c_{t,M}]^\top$ that maximizes GP_t, the value of the function described in Eq.(5.31) at time frame t. We then have:

$$\boldsymbol{c}_t^* = \arg\max GP_t(\boldsymbol{c}_t) \tag{6.5}$$

$$GP_t(\boldsymbol{c}_t) = \frac{100}{N_f} \sum_{f=1}^{N_f} \mathcal{L}(y_{t,f}^{sp}(\boldsymbol{c}_t) - y_{t,f}^{ns}) \tag{6.6}$$

See here that in contrast to the cost function of the GP-based cepstral coefficient proposed in the previous chapter, this cost function only depends on the GP measure.

6.3.2 Steepest descent solution

As the function we are maximizing is not necessarily convex with respect to the Mel cepstral coefficients, we use a Steepest Descent method to solve the optimization. The update equation is:

$$\boldsymbol{c}_t^{(i+1)} = \boldsymbol{c}_t^{(i)} + \mu \nabla GP_t^{(i)} \tag{6.7}$$

where $\Delta c^{(i)}$ is the Mel cepstral coefficient increment in iteration i, $\nabla GP_t^{(i)}(\boldsymbol{c}_t)$ is the gradient of the function defined in Eq.(6.6) with regards to Mel cepstral coefficients in iteration i and μ is the stepsize.

From now on we will drop the iteration index (i). As we showed in the previous chapter, we can find the gradient vector as follows:

$$\nabla GP_t = \frac{\partial GP_t}{\partial \boldsymbol{c}_t} = \frac{100}{N_f} \sum_{f=1}^{N_f} \frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} \tag{6.8}$$

The first term in this summation, see Eq.(5.44), is given by:

$$\frac{\partial \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})}{\partial y_{t,f}^{sp}} = \eta\, \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns}) \left[1 - \mathcal{L}(y_{t,f}^{sp} - y_{t,f}^{ns})\right] \tag{6.9}$$

The second term in this summation, see Eq.(5.45), is given by:

$$\frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{c}_t} = \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} \frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{h}_t} \tag{6.10}$$

where:

$$\boldsymbol{H}_{c_t} \equiv \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{c}_t} = \begin{bmatrix} \frac{\partial |H_t(\omega_1)|}{\partial c_{t,1}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,1}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,1}} \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,2}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,2}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,2}} \\ \vdots & & & \\ \frac{\partial |H_t(\omega_1)|}{\partial c_{t,M}} & \frac{\partial |H_t(\omega_2)|}{\partial c_{t,M}} & \cdots & \frac{\partial |H_t(\omega_N)|}{\partial c_{t,M}} \end{bmatrix}$$

When the spectrum is modelled by Mel cepstral coefficients as in Eq.(6.4) the elements of this matrix are:

$$\{\boldsymbol{H}_{c_t}\}_{m,k} = \frac{\partial |H_t(\omega_k)|}{\partial c_{t,m}} = |H_t(\omega_k)| \cos(m\,\tilde{\omega}_k) \tag{6.11}$$

Following the same derivation as seen in the last chapter Section 6.3, the second term of Eq.(6.10) depends on the definition of the STEP approximation presented in the last chapter , see Eq.(5.33), and it is then given by:

$$\frac{\partial y_{t,f}^{sp}}{\partial \boldsymbol{h}_t} = \frac{1}{N}\boldsymbol{G}_f\left(2\boldsymbol{\Gamma}_N \textcircled{N} \boldsymbol{G}_f\boldsymbol{h}_t\right)\boldsymbol{S}\boldsymbol{b} \tag{6.12}$$

where $\boldsymbol{\Gamma}_N$ is the identity matrix of dimension N, $\boldsymbol{S}$ is an NxN diagonal matrix, whose diagonal contains the frequency response of the smoothing filter of the GP approximation and $\boldsymbol{b}$ a Nx1 vector containing the coefficients of the average filter used for the GP approximation. The operation $(\boldsymbol{\Gamma}_N \textcircled{N} \boldsymbol{G}_f\boldsymbol{h}_t)$ defines a matrix NxN of the following form:

$$\begin{bmatrix} \boldsymbol{e}_1 \textcircled{N} (\boldsymbol{G}_f\boldsymbol{h}_t)^\top \\ \boldsymbol{e}_2 \textcircled{N} (\boldsymbol{G}_f\boldsymbol{h}_t)^\top \\ \vdots \\ \boldsymbol{e}_N \textcircled{N} (\boldsymbol{G}_f\boldsymbol{h}_t)^\top \end{bmatrix}$$

where $\boldsymbol{e}_n$ is the n-th column of the identity matrix $\boldsymbol{\Gamma}_N$. The gradient vector is given by:

$$\nabla GP_t = \frac{100}{N_f N}\sum_{f=1}^{N_f}\eta\,\mathcal{L}(y_{t,f}^{sp}-y_{t,f}^{ns})\left[1-\mathcal{L}(y_{t,f}^{sp}-y_{t,f}^{ns})\right]\cdot \boldsymbol{H}_{c_t}\boldsymbol{G}_f\left(2\boldsymbol{\Gamma}_N \textcircled{N} \boldsymbol{G}_f\boldsymbol{h}_t\right)\boldsymbol{S}\boldsymbol{b} \tag{6.13}$$

Note here that the only difference between the GP gradient calculated here and the one calculated for the GP-based cepstral coefficient extraction is the fact that vector $\boldsymbol{h}_t$ and matrix $\boldsymbol{H}_{c_t}$ are defined by Mel cesptral coefficients.

6.3.3 Energy normalization

A derivation similar to what has been showed in the previous chapter in Section 5.5.3 is shown here for finding the steepest descent solution that includes an energy normalization implicit constraint.

We do not wish to modify the energy of the speech signal:

$$\psi = \sum_{n=0}^{N-1}|s(n)|^2 \tag{6.14}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}|S(e^{j\omega})|^2\,d\omega \tag{6.15}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}|H(e^{j\omega})E(e^{j\omega})|^2\,d\omega \tag{6.16}$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}|H(e^{j\omega})|^2\,d\omega \tag{6.17}$$

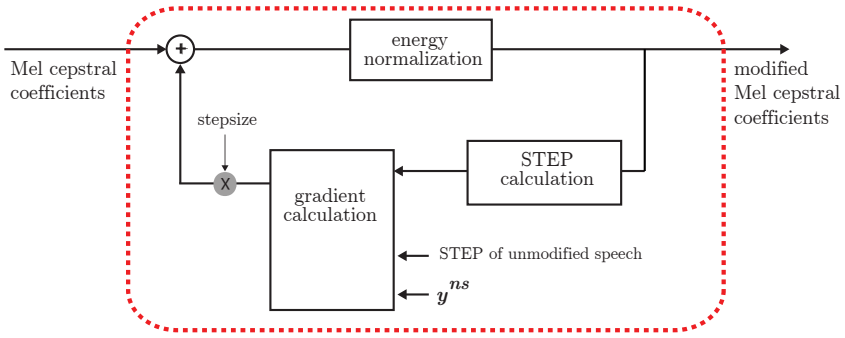


Figure 6.1: GP-based Mel cepstral coefficient modification using steepest descent with energy normalization.

where N is the size of the time window frame, $S(e^{j\omega})$ is the discrete time Fourier transform of time signal $s(n)$, $H(e^{j\omega})$ and $E(e^{j\omega})$ the frequency representation of the excitation signal. The equality of the last equation arises from the assumption that $E(e^{j\omega})$ is constant over the frequency domain, see Section 5.5.3 for further explanation.

Finding a closed solution for a non-convex constrained optimization problem is a hard task. One possible solution is to solve it with steepest descent by adding an energy normalization step at each iteration and changing the gradient accordingly. Fig. 6.1 shows this solution.

The normalization operation transforms a set of Mel cepstral coefficients c_m that model the spectrum $|H(e^{j\omega})|$ with overall energy ψ, into parameters c'_m that model a spectrum $|H'(e^{j\omega})|$ with overall energy equal to ψ' in the following way:

$$|H'(e^{j\omega})| = \frac{|H(e^{j\omega})|}{\sqrt{\frac{\psi}{\psi'}}} = \frac{\exp\sum_{m=0}^{M} c_m \cos(m\tilde{\omega})}{\exp\left[\log\sqrt{\frac{\psi}{\psi'}}\right]} \tag{6.18}$$

$$= \exp\left[\left(\sum_{m=0}^{M} c_m \cos(m\tilde{\omega})\right) - 0.5\log\left(\frac{\psi}{\psi'}\right)\right] \tag{6.19}$$

$$= \exp\sum_{m=0}^{M} c'_m \cos(m\tilde{\omega}) \tag{6.20}$$

The energy-normalized Mel cepstral coefficients c'_m are then given by:

$$c'_m = \begin{cases} c_0 - 0.5\log\left(\frac{\psi}{\psi'}\right) & m = 0 \\ c_m & m \neq 0 \end{cases} \tag{6.21}$$

Only the c_0 coefficient changes, so we can write the energy normalized magnitude spectrum as:

$$|H'(e^{j\omega})| = |K'||D(e^{j\omega})| \tag{6.22}$$

where $K' = \exp(c'_0)$ and $D(e^{j\omega}) = \exp \sum_{m=1}^{M} c_m e^{-jm\tilde{\omega}}$.
With the energy normalization operation, the derivative in Eq.(6.11) becomes:

$$\frac{\partial |H'(\omega_k)|}{\partial c_m} = \frac{\partial |K'||D(\omega_k)|}{\partial c_m} \tag{6.23}$$

$$= \frac{\partial |K'|}{\partial c_m}|D(\omega_k)| + |K'|\frac{\partial |D(\omega_k)|}{\partial c_m} \tag{6.24}$$

$$= |K'|\frac{\partial c'_0}{\partial c_m}|D(\omega_k)| + |K'||D(\omega_k)|\cos(m\tilde{\omega}_k) \tag{6.25}$$

$$= |H'(\omega_k)|\frac{\partial c'_0}{\partial c_m} + |H'(\omega_k)|\cos(m\tilde{\omega}_k) \tag{6.26}$$

$$= |H'(\omega_k)|\Big(\frac{\partial c'_0}{\partial c_m} + \cos(m\tilde{\omega}_k)\Big) \tag{6.27}$$

The derivative term in the previous equation is given by:

$$\frac{\partial c'_0}{\partial c_m} = \frac{\partial c_0}{\partial c_m} - 0.5\,\frac{\psi'}{\psi}\,\frac{1}{\psi'}\,\frac{\partial \psi}{\partial c_m} \tag{6.28}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi}\sum_{l=1}^{N}|H(\omega_l)|^2\cos(m\tilde{\omega}_l) \tag{6.29}$$

$$= \frac{\partial c_0}{\partial c_m} - \frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\tilde{\omega}_l) \tag{6.30}$$

$$\frac{\partial c'_0}{\partial c_m} = \begin{cases} 0.0 & m = 0 \\ -\frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\tilde{\omega}_l) & m \neq 0 \end{cases} \tag{6.31}$$

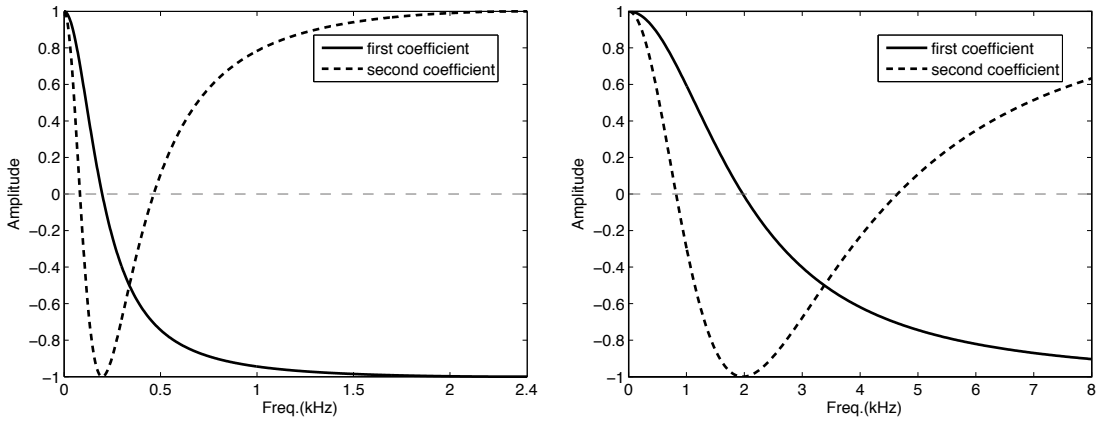
The derivative in Eq.(6.11) then becomes:

$$\frac{\partial |H'(\omega_k)|}{\partial c_m} = \begin{cases} |H'(\omega_k)| & m = 0 \\ |H'(\omega_k)|\Big(\cos(m\tilde{\omega}_k) - \frac{1}{\psi'}\sum_{l=1}^{N}|H'(\omega_l)|^2\cos(m\tilde{\omega}_l)\Big) & m \neq 0 \end{cases} \tag{6.32}$$

As it was for the previous chapter, the frame-by-frame energy normalization implies that no energy is reallocated across time and the modifications are therefore limited by the energy initially available in each analysis window. It is possible to prove that is unnecessary to update the first Mel cepstral coefficient c_0 in each iteration since the normalization operation updates it to a certain value regardless of the additional Δc_0 term.

6.3.4 Distortion control

An audibility-based measure like GP predicts the effect of additive noise by comparing the levels of speech and noise. If speech is modified in a way that it creates many



(a) full frequency band (b) narrow frequency band

Figure 6.2: Cosines associeted with the first c_1 and second c_2 Mel cepstral coefficient in the (a) full frequency band and (b) the narrow frequency band.

glimpses but also generates distortions, the GP measure will increase independent of the distortion. This happens because the GP does not require any reference of undistorted speech signal: it assumes that speech has not been modified. An issue we face then when using the GP measure on its own as an optimization criterion is the need to limit the distortions caused by the modifications. For instance, if the spectral envelope is modified to maximize the number of glimpses without any additional constraint this would result in a spectral envelope just above the noise spectrum, i.e. speech would be shaped by the noise. Recent research on improving the GP measure to account for speech that has been modified is described in Tang et al. (2013). Tang et al. (2013) proposes to weight the time-frequency bins defined by the STEP representation with the cross-correlation of the temporal envelopes of clean unmodified and noisy modified speech. Our work was however based on the original measure (Cooke, 2006).

To define the audible distortion, we use the Euclidian distance between the STEP representations of modified and unmodified speech. Including this as an explicit constraint is hard because the constraint is non linear to the variable we are optimizing upon, in addition to non-convexity of the problem. Instead, we use it as a stopping criterion.

We also hypothesize that limiting the frequency resolution of the modifications should generate fewer distortions. This is implemented simply by setting the gradient vector for higher dimensions to zero, and so the method modifies only the first few Mel cepstral coefficients, which represent the coarse properties of the spectrum.

To illustrate the effect that each coefficient has on the composition of the magnitude

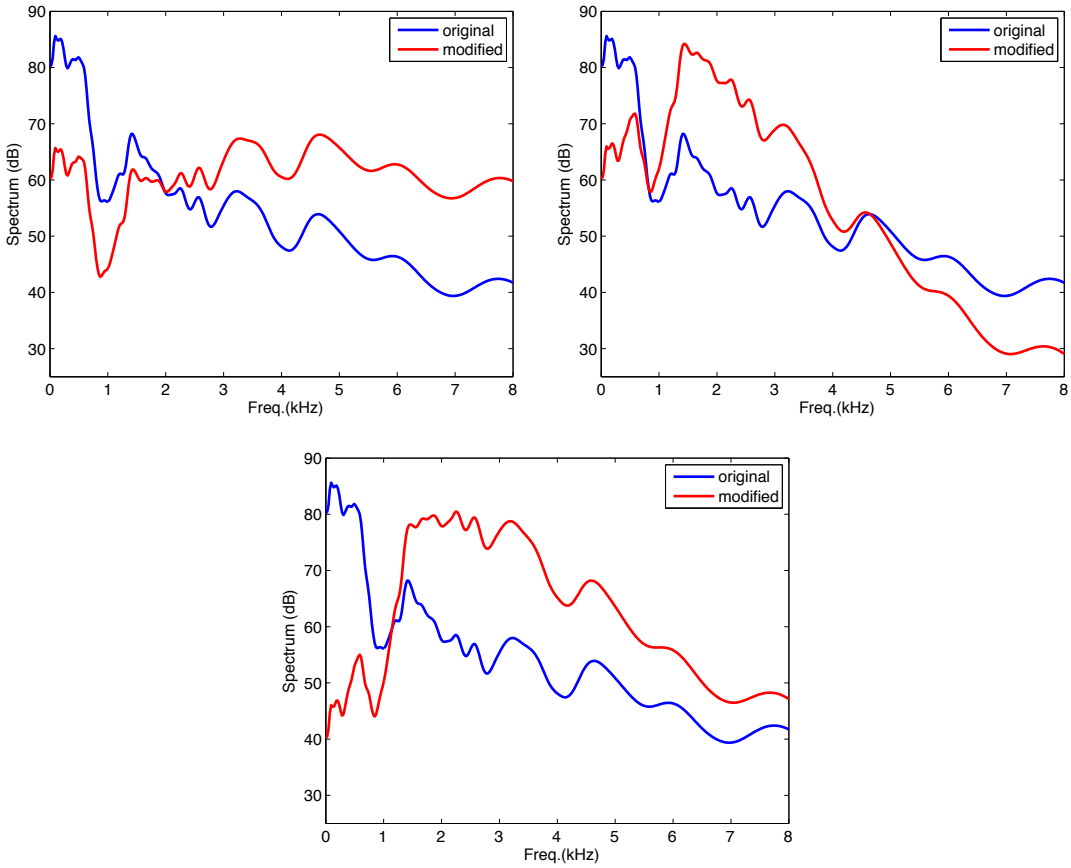


Figure 6.3: Original and modified spectral envelope. Modification here refers to decreasing the value of the first coefficient c_1 (top left), the second c_2 (top right) and both first c_1 and second c_2 Mel cepstral coefficients.

spectrum we present, in Fig. 6.2, the cosines in the summation that define the log-spectrum whose amplitude are the first and second Mel cepstral coefficients. The left hand figure shows the cosine covering a frequency range of 0-24 kHz and the figure on the right zooms into the narrower band of 0-8 kHz which will help us understand the effect of our results when we extract Mel cepstral coefficients from a signal sampled at 48 kHz and then downsample the signal to 16 kHz.

In Fig. 6.3, we can see the effect of changing the first, second and both first and second Mel cepstral coefficients on the long term average spectrum of speech (LTAS). To create these figures, we subtracted a constant value from the first, second and both first and second coefficients defining a speech segment and calculated the LTAS of the unmodified (original) and the modified segment. We can see that changing the first coefficient allows for energy reallocation across two different frequency regions. If the cepstral coefficients were defined for a linear scaled spectrum then these regions

would divide the whole frequency range in half, as the first coefficient corresponds to the amplitude of the first cosine defining the log spectrum. Since these are Mel cepstral coefficients however, these two regions span the same frequency range on the Mel scale. This means that changing the first Mel cepstral coefficient allows us to reallocate energy from quite low frequency components, that on average have higher energy levels, to higher frequency regions. Changing the second Mel cepstral coefficient defines three different regions as we are now changing the amplitude of the second cosine. Although we do not present a proof here, we expect that changing both first and second coefficients allows for a more independent control of the region boundaries, the number of regions and their gains.

Algorithm using Steepest Descent and energy normalization

Given the Mel cepstral coefficient set $\boldsymbol{c}_t$ and the noise waveform, for time frame t:

- Calculate internal representation of noise: $y_t^{ns} = \left[y_{t,1}^{ns} \ \dots \ y_{t,N_f}^{ns} \right]^\top$
- Calculate the overall energy ψ' of the spectrum modelled by Mel cepstral coefficients
- Optimization loop:
  - Normalize spectrum so that overall energy is $\psi' \rightarrow \boldsymbol{h}'_t$ and $\boldsymbol{c}'_t$
  - Given the new spectral envelope $\boldsymbol{h}'_t$ calculate the speech internal representation $y_t^{sp} = \left[y_{t,1}^{sp} \ \dots \ y_{t,N_f}^{sp} \right]^\top$
  - Calculate gradient vector ∇GP_t
  - Update $\boldsymbol{c}_t$ and calculate $\boldsymbol{h}_t$
  - If converges or distortion is above threshold then stop.

6.4 Evaluation

In this section, we investigate whether the proposed Mel cepstral modification can increase subjective scores of intelligibility. First however we present the details of how the synthesis models were built and the parameters that we set for the proposed modification. We want to evaluate the idea of restricting the frequency resolution of the modifications by updating only the first few Mel cepstral coefficients. For that we cre-

Voice	Adaptation	Mel cepstral coefficient modification
N	-	-
N-M59	-	all coefficients
N-M10	-	first 10 coefficients
N-M2	-	first 2 coefficients
N-L	only spectral parameters	-
L	all parameters	-
L-E	all parameters extrapolated	-
L-E-M2	all parameters extrapolated	first 2 coefficients

Table 6.1: Voices built for the evaluation.

ate a range of TTS voices by changing the number of coefficients to be modified. As an additional baseline comparison we include a TTS voice trained with Lombard speech – speech produced in noise – of the same speaker. We present a detailed acoustic analysis of the modified synthetic speech signal and then describe how we designed the listening experiments and finally the results of the experiment.

6.4.1 Voice building

We used two different datasets recorded by the same British male speaker *Nick*: normal (plain, read-text) speech data and Lombard speech. The Lombard speech was recorded by Cooke et al. (2012) while speech-modulated noise (modulated by the speech from a different male speaker (Dreschler et al., 2001)) was played over headphones at a absolute value of 84 dBA.

Table 6.1 presents the eight different voices we built for this evaluation. The baseline unmodified voice N was created from a high quality average voice model adapted to 2803 sentences of the normal speech database (three hours of material). The adaptation technique used to create the adapted voices of this work is the CSMAPLR-MAP adaptation described in (Yamagishi et al., 2009). The average voice model was built with female British speaker data and it provided smaller likelihood than a male model, more details on how the model was built can be found in Dall et al. (2012). Building a speaker-dependent voice was not possible because the normal speech dataset was not sufficiently phonetically balanced. The modified voices N-M59, N-M10 and N-M2 were created from voice N by modifying all, just the first ten (c_1 until c_{10}), or just the

first two (c_1 and c_2) Mel cepstral coefficients using the proposed method, as described in the previous section.

We built the other set of voices N-L, L, L-E and L-E-M2 using the Lombard speech portion of the database in addition. Lombard voice L was built by further adapting all parameters (duration, excitation, spectral) of voice N using 780 sentences from the Lombard speech dataset (53 minutes). The reason for not building a voice only with the Lombard dataset was again the lack of phonetic balance in the dataset. Voice N-L was also created from voice N by adapting this time only the Mel cepstral coefficients (i.e., spectral model parameters) to the Lombard data. Voices L-E and L-E-M2 are versions of voice L where we extrapolated the adaptation in all dimensions at an extrapolation factor of 1.2 for Mel cepstral coefficients and 1.35 for duration (voice L-E), and then further modified the two first Mel cepstral using the proposed method (voice L-E-M2). The extrapolation factors used were found empirically as the maximum extrapolation factor that does not generate audible artefacts.

We trained and adapted the models using the described data sampled at a rate of 48 kHz. We extracted the following acoustic features: 59 Mel cepstral coefficients (α=0.77), Mel scale F0 and 25 aperiodicity band energies extracted using STRAIGHT (Kawahara et al., 1999). We used a hidden semi-Markov model as the acoustic model. The observation vectors for the spectral and excitation parameters contained static, delta and delta-delta values, with one stream for the spectrum, three streams for F0 and one for the band-limited aperiodicity. We applied the Global Variance method (Toda and Tokuda, 2007) to compensate for the over smoothing effect caused by the statistical nature of the acoustical modelling. The labels used to train and generate the test sentences were built using the pronunciation lexicon Combilex (Richmond et al., 2010).

For the GP-based Mel cepstral modifications, we set the following values for the STEP calculation: 55 Gammatone filters with centre frequencies covering the range of 50-7500 Hz (because the noise signal used for testing was sampled at 16 kHz, and so the audio bandwidth was 8 kHz), 8 ms of temporal integration time for the smoothing filter and frame length and period of 30 and 10ms. For the steepest descent optimization we used a normalized step size defined at each iteration i as $\mu^{(i)} = \mu / ||\nabla GP_t^{(i)}||$ (where μ=0.4 for N-M59 and μ=0.8 for N-M10 and N-M2). As stopping criteria we use both error convergence and a maximum threshold set to 10 % of relative increase in distortion. We define distortion here as the Euclidian distance between the original and the modified STEP representation of speech. After synthesizing, the speech wave-

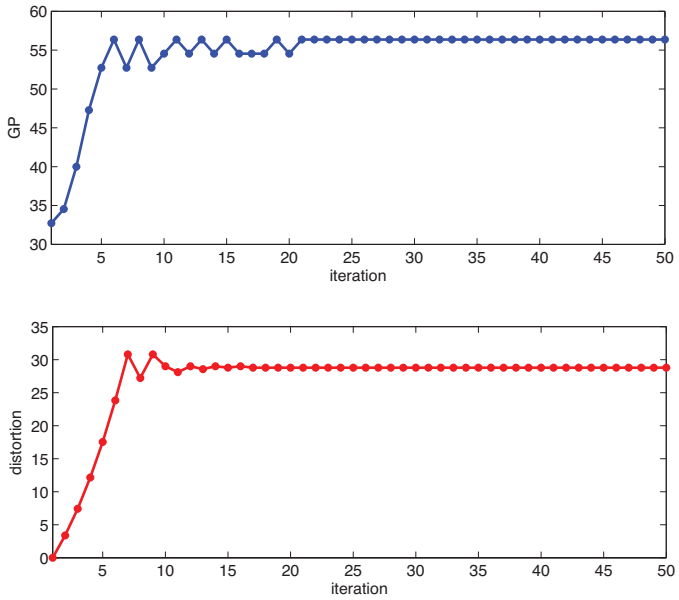


Figure 6.4: Convergence of GP (top) and distortion (bottom) for a certain single time frame. Distortion is measured as the percentage increase in the Euclidian distance between the STEP representation of original and modified spectrum. The stopping criterion was not applied to illustrate the convergence.

form was downsampled to 16 kHz. This was necessary because the noise signals were produced at this lower sampling rate.

6.4.2 Convergence analysis

Fig. 6.5 shows the convergence of the GP and distortion values. We can see that, as GP increases, distortion also increases as expected, and that the algorithm is well-behaved (i.e., it converges to a stable value within a reasonable number of iterations). The algorithm is frame-based, meaning that the stopping criteria are applied on a per-frame basis. For individual frames, the convergence presented in Fig. 6.4 is somewhat less smooth-looking than that illustrated in the Fig. 6.5. On average, five iterations are sufficient to meet one or other of the stopping criteria for each frame, and more often than not it is the distortion criterion that is met first.

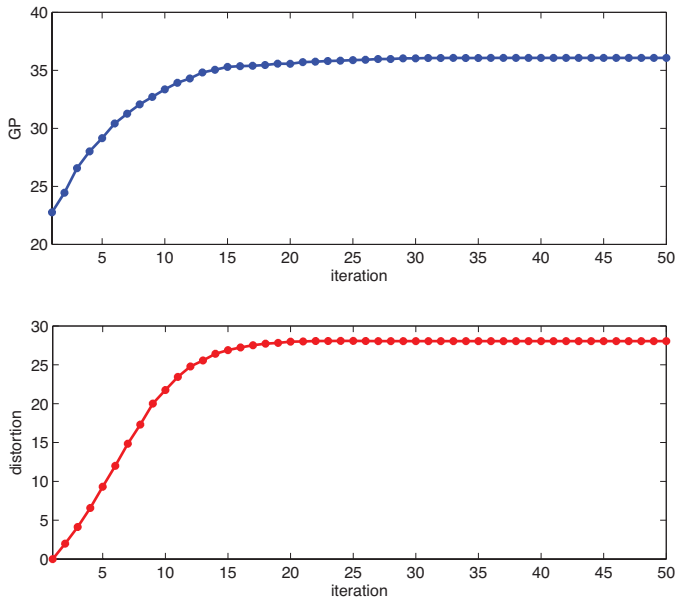


Figure 6.5: Convergence of GP (top) and distortion (bottom), averaged over all frames of one sentence. Distortion is measured as the percentage increase in the Euclidian distance between the STEP representation of original and modified spectrum. The stopping criterion was not applied to illustrate the convergence.

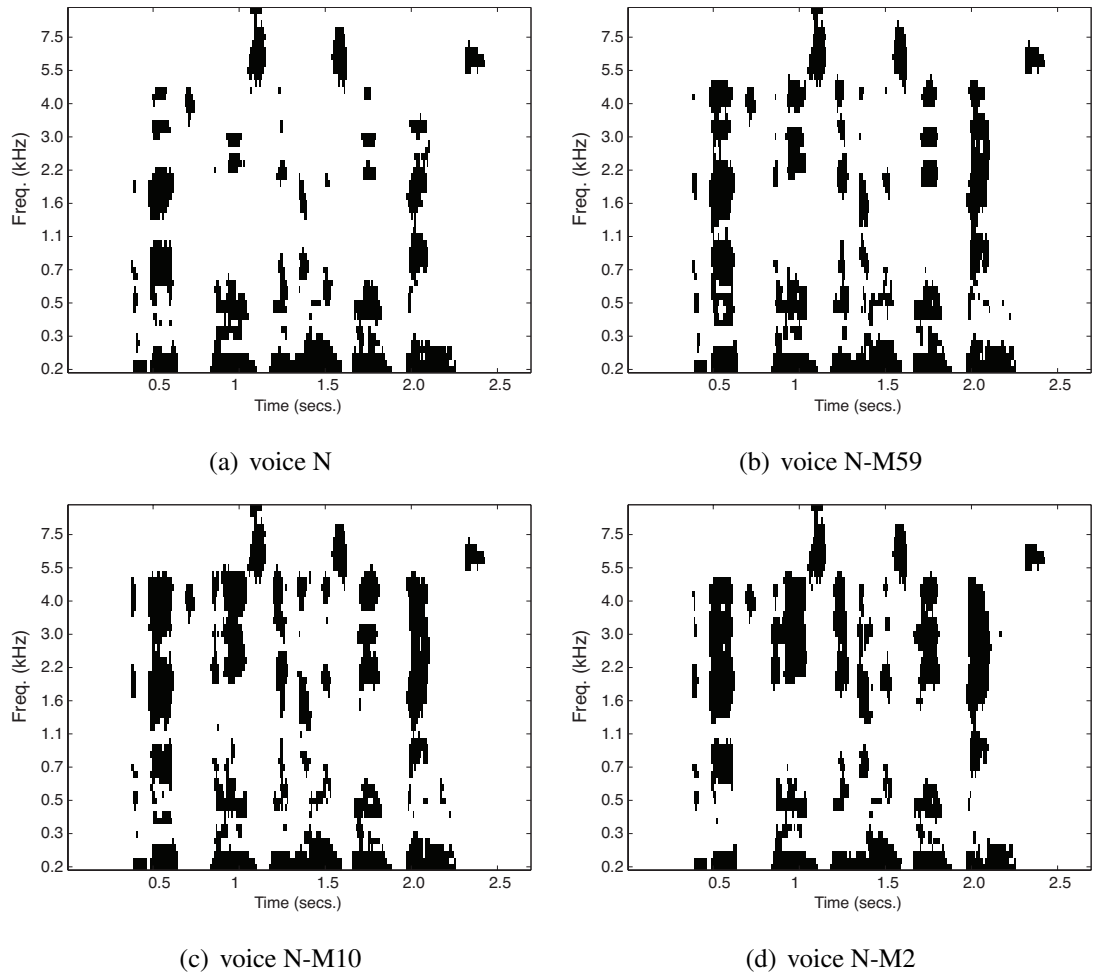
6.4.3 Acoustic analysis

In this section, we examine the impact of the modifications on the whole set of sentences and at sentence and phone class level. The results are shown in terms of spectral tilt, GP values and spectral gains, calculated using the long term average spectrum (LTAS).

First, we present a broad analysis across the whole set of sentences used in the listening experiment. Table 6.2 shows the average duration of speech and pauses, average F_0 and average spectral tilt across all sentences used in the listening test for the normal (N), modified (N-M2) and Lombard (L) voices. We can see that, as expected, the Lombard voice produces sentences with longer duration and longer pauses, greatly increased F_0 mean and flattening of the spectral tilt. The spectral tilt reflects changes in both spectral envelope and excitation signal. The modified voice N-M2 also presents a flatter spectral tilt, though not to the same extent as the Lombard voice.

voice	duration (secs.)	pauses (secs.)	F_0 mean (Hz)	spectral tilt (dB/oct.)
N	2.11	0.16	104.5	-2.24
N-M2				-1.88
L	2.80	0.19	145.0	-1.70

Table 6.2: Acoustic properties observed in normal N, modified N-M2 and Lombard L voices calculated at a sentence level and averaged across the whole set of 110 sentences.



(a) voice N (b) voice N-M59

(c) voice N-M10 (d) voice N-M2

Figure 6.6: Glimpses detected on the STEP time-frequency representation in speech-shaped noise at a SNR of -4 dB (in black) for a sentence generated by (a) unmodified voice N and modified voices (b) N-M59 (c) N-M10 and (d) N-M2

For a more detailed inspection (sentence-level) of the proposed method in operation, Fig. 6.6 shows the glimpses (in black) detected in the presence of speech-shaped noise at -4 dB SNR for (from left to right) a sentence generated by the unmodified voice N and the modified voices N-M59, N-M10 and N-M2. The glimpses are shown here in the STEP frequency domain across different time frames, the frequency axis is linearly spaced in the ERB frequency scale as defined by the Gammatone filter bank used to extract the STEP representation. We can see that the glimpsed regions become larger and that new glimpses start to appear when we modify all, just the first ten and the first two Mel cepstral coefficients. We also see that when we modify fewer coefficients, the new glimpses tend to be in more coherent regions, creating larger glimpses rather than scattered small glimpses. This is an expected and desired result of modifying only those coefficients that define the coarse shape of the log magnitude spectrum.

Fig. 6.7 shows the GP value for each frame as defined in Eq.(6.6) for the same sentence as shown in Fig. 6.6, generated by the unmodified voice N and the modified voice N-M2, together with the segmented phone description as defined by the combilex phoneset for that particular sentence *"The birch canoe slid on the smooth planks"*.

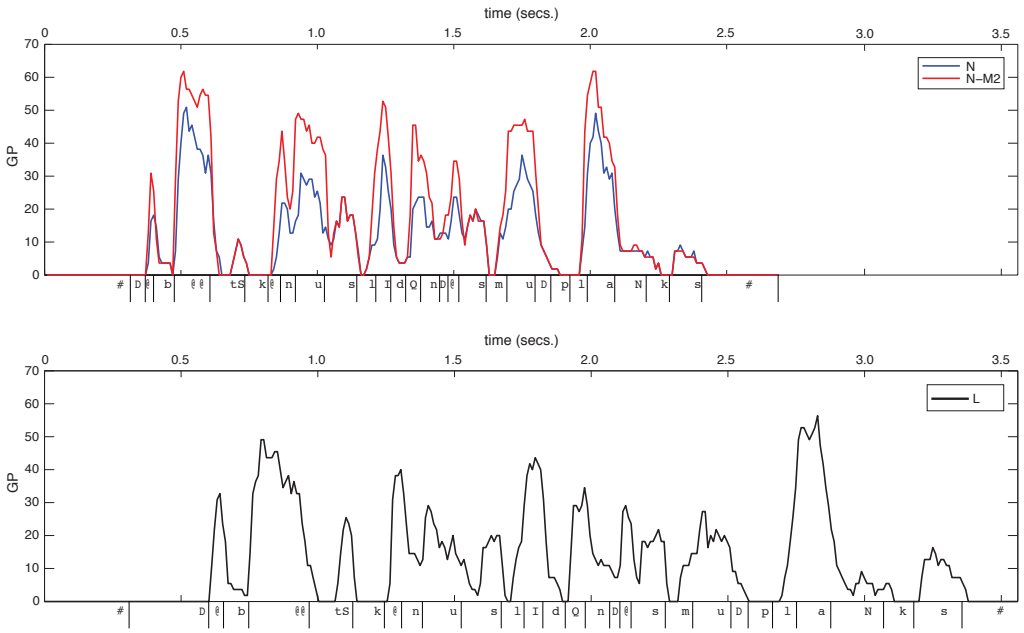


Figure 6.7: The GP measure across the different frames of the sentence *"The birch canoe slid on the smooth planks"* generated by the original unmodified voice N versus the modified voice N-M2 (top) and the Lombard adapted voice L (bottom), in the presence of speech-shaped noise at -4 dB. The horizontal axis gives the phone segmentation in the Combilex phoneset.

We observe in Fig. 6.7 that, although the number of glimpses on average increases, the increase in glimpses differs between segments. Since the noise that was driving this modification is stationary, this variation comes from the speech signal itself: the different spectral shapes of the various phonetic units will result in fewer or greater numbers of glimpses. In this example sentence, the number of glimpses hardly increases in fricatives and stops, whereas the most substantial increases happen in vowels and nasals. This does not mean that fricatives and stops are not being modified though, but does mean that the proposed method fails to create more glimpses of them for the listener. Although we are not aiming to recreate the Lombard effect, we present in the bottom plot of Fig. 6.7 the GP values calculated using the voice L. Compared to the GP gains obtained by voice N-L over voice N, the voice L has smaller GP gains during vowels while fricatives' GP values are slightly higher.

To find which frequency regions are boosted and which are attenuated, we compute the spectral gain in (dB). The spectral gain is a measure that captures the gain over a reference LTAS curve. It is calculated as the difference between the LTAS of the signal and the LTAS of the reference signal, in our case the voice N. The LTAS is calculated as the averaged power spectral density calculated using hamming windows of 10 ms length and 50 % overlap. This averaged representation is then presented in (dB) by means of $10 * \log_{10}$ operation. We computed the spectral gain of voice N-M2 over the original unmodified voice N, averaged across all test sentences, for speech-shaped noise at -4 dB. Fig. 6.8 shows the overall pattern of spectral gain at a sentence level.

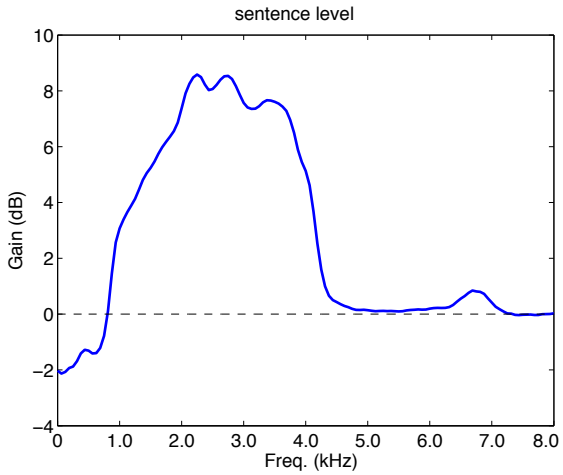


Figure 6.8: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) at a sentence level and averaged across the set of sentences used in the listening test.

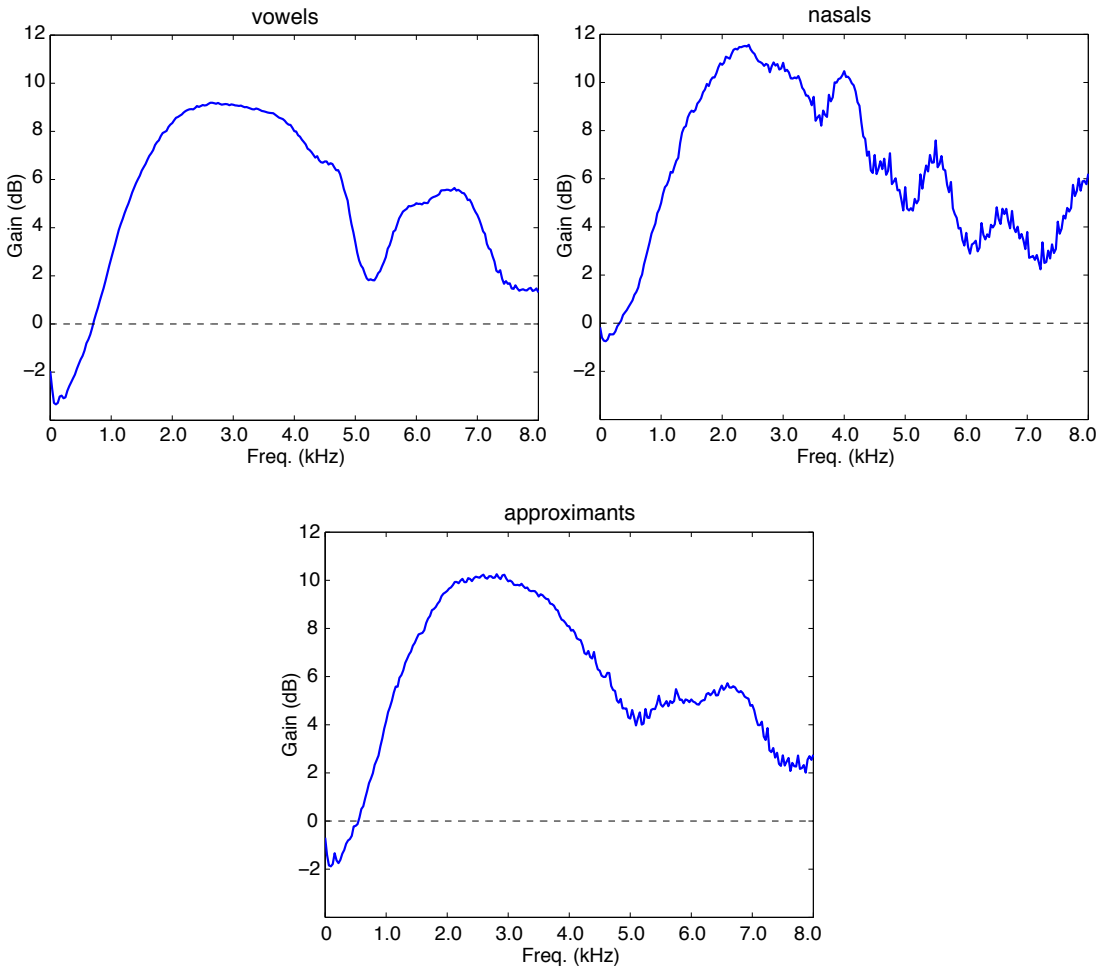


Figure 6.9: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) averaged across vowels (top left), nasals (top right) and approximants (bottom).

From Fig. 6.8, we observe that, compared to voice N, voice N-M2 exhibits enhanced energy in the region of 1-4 kHz and attenuated energy below 1 kHz.

To observe how different segments of speech change, we calculated the gains curves for different phonetic classes using the state boundaries to select the units. Figs. 6.9 and 6.10 present the gain calculated for different phonetic classes averaged over all tokens of that class in the test set. One clear observation we can make when comparing the gains for specific phone classes as displayed in Figs. 6.9 and 6.10 is that the curves as well as the gain values vary substantially across different phonetic classes. In the first group (vowels, nasals and approximants) the gains are at least five times larger than those obtained for the second group (fricatives, affricates and stops). This is a consequence of the shapes and values of the unmodified speech LTAS

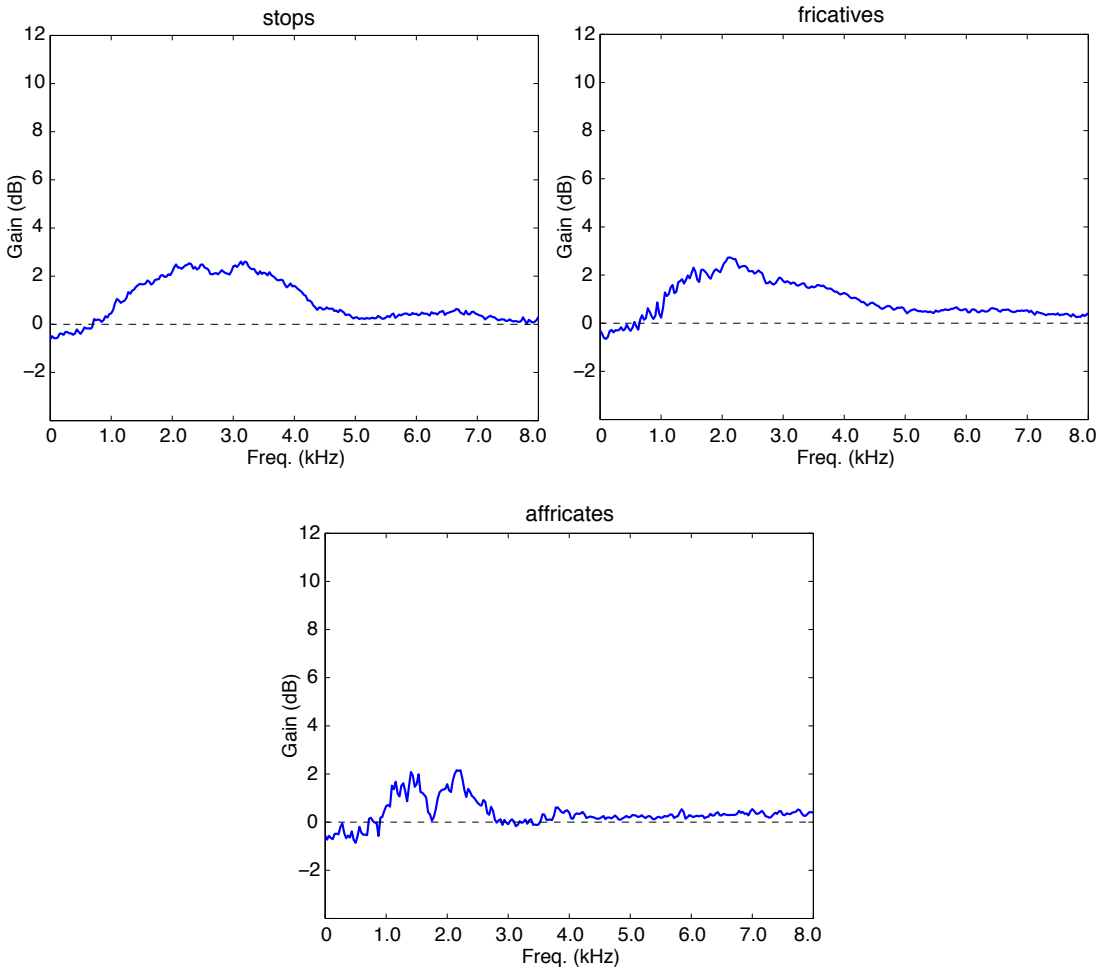


Figure 6.10: Gain in (dB) of the LTAS of voice N-M2 over the LTAS of unmodified voice N calculated (for speech-shaped noise) averaged across stops (top left), fricatives (top right) and affricates (bottom).

for these classes. We present the LTAS curves of unmodified speech segments averaged across these classes in the Figs. 6.11 and 6.12 as well as the noise LTAS curve, speech-shaped noise at -4 dB SNR as a reference. As we can see in Figs. 6.11 and 6.12 the first group of phonetic units LTAS curves reach levels of up to 40 dB at lower frequencies and are steeper compared to the second group whose LTAS curve values are far from reaching the noise LTAS and are also quite flat. We can see that there was probably not enough "energy budget" in those phonetic units to make the substantial modifications that need to be made in order to increase the number of glimpse regions in the spectrum, which resulted in lower LTAS gains and very few gains in terms of GP as we observed in the Fig. 6.7.

From the gain curves of the first group displayed in Fig. 6.9 we can see a similar

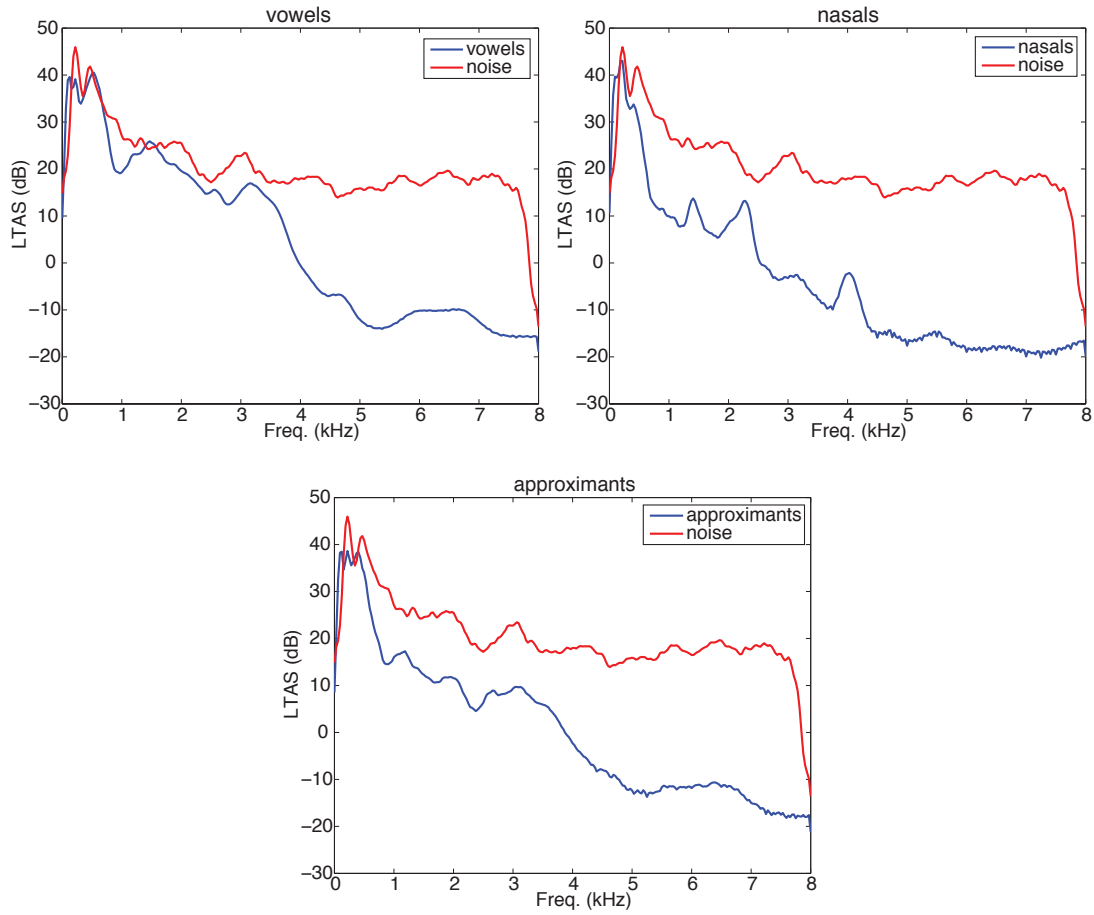


Figure 6.11: Long term average spectrum (LTAS) of speech shaped noise and unmodified speech average across vowels (top left), nasals (top right) and approximants (bottom).

pattern across vowels, nasals and approximants: a large enhancement varying from 8 to 12 dB in the frequency region between approximatively 800 Hz (this number varies across the different classes) and 5 kHz as well as an apparent attenuation of around 2 dB for the lower frequency region. For both vowels and approximants, we also see a clear gain region between 5-8 kHz that is separated by a gain reduction at approximately 5 kHz. The shapes of these gain curves follow the shape of the LTAS of these phonetic classes, for instance we can see a bump from 5-8kHz in the vowels and approximants. The nasals are the units that are most strongly enhanced reaching a maximum of 12 dB gain which can be explained by the fact that they seem to be highly energetic with an even less flat spectrum than the other sounds.

A similar trend for vowels, nasals and liquids can be seen in a study performed on Lombard speech of 5 male Spanish native speakers (Castellanos et al., 1996). The

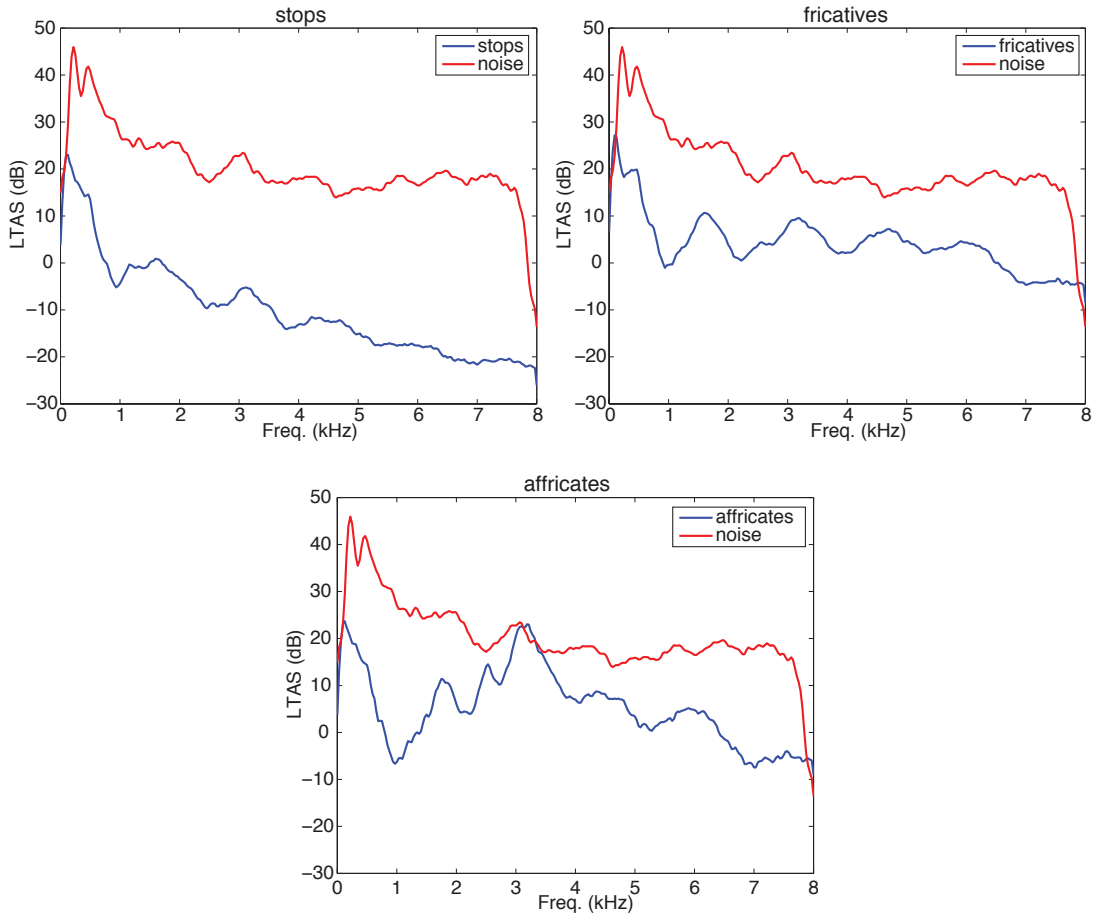


Figure 6.12: Long term average spectrum (LTAS) of speech shaped noise and unmodified speech average across stops (top left), fricatives (top right) and affricates (bottom).

GRID Lombard corpus composed of 8 English native speakers also shows a spectral gain with similar bi-modal characteristics – a peak in the formant region and a slight peak in the 6 – 7kHz region – as the sentence spectral gain presented in Fig. 6.8 (Godoy and Stylianou, 2012). Spectral gains of the Lombard database that was used in this work also present a bi-modal nature but the high frequency mode is more boosted than the formant region, see Appendix B for the curves.

The gains obtained for the other class (stops, fricative and affricates) are, as previously stated, much smaller. For both stops and fricatives an average maximum of 2 dB increase was found and the region most enhanced is between 1-5 kHz as seen for the other group. The affricates show even lower gains and narrow enhanced regions between 1-3 kHz with a valley around 2 kHz.

On average across different phonetic units in the same sentence we show in Fig. 6.13 the long term average spectrum of the normal (N), modified (N-M2) and Lombard (L)

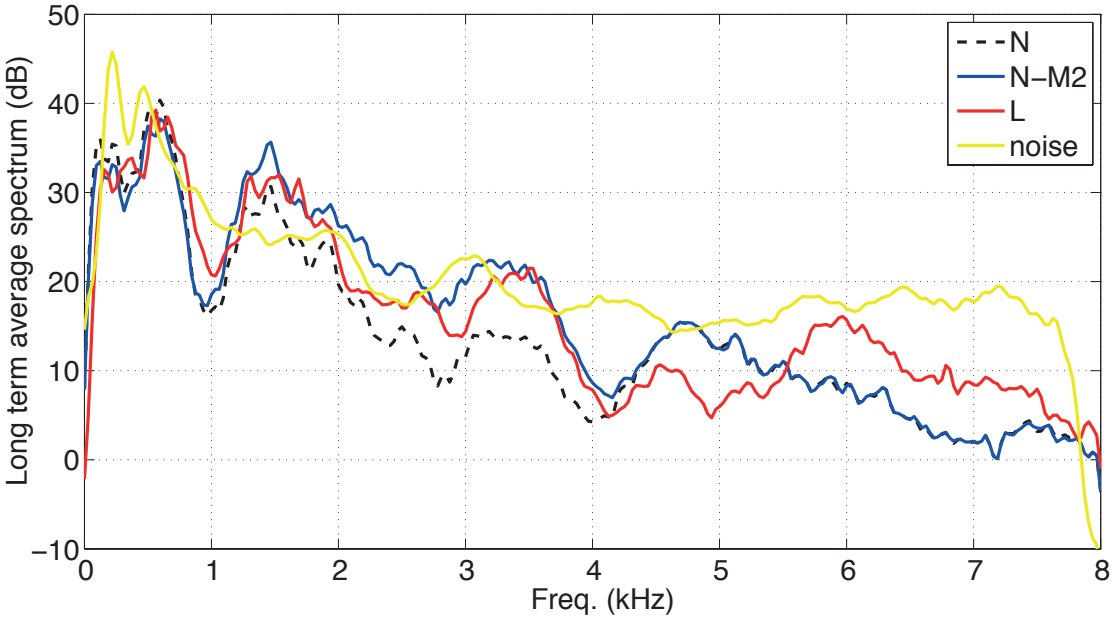


Figure 6.13: Long term average spectrum of the normal N, normal modified N-M2 and Lombard L voices for speech-shaped noise.

voices, for the case of speech-shaped noise. Compared to voice N, voice N-M2 exhibits enhanced energy in the frequency region of 1-4 kHz and attenuated below 1 kHz. Voice L shows enhancement and attenuation in the same regions as N-M2, although these changes are not as pronounced, attenuation is also seen between 4-5.5 kHz and enhancement at frequencies above this.

6.4.4 Listening experiments

We mixed the eight different synthetic voices with two noises: speech-shaped noise and speech from a single competing female talker. For intelligibility testing, it is important to avoid floor or ceiling effects on word accuracy rate. Therefore, in order to obtain intelligibility scores in similar ranges for each noise, we mixed them at differing SNRs: -4 dB for speech-shaped noise and -14 dB for the competing speaker. As in the previous chapter, no other energy normalization had to be performed to guarantee that the energy level of the sentence was not modified, since the GP-based modification proposed here does not modify the energy of the signal.

32 native English speakers listened to the noisy samples over headphones in sound-isolated booths. Each participant typed in what he or she heard for a total of six different sentences per condition, i.e., voice and noise type (16 conditions). Each sentence could only be played once and the same sentence was never played again in the same

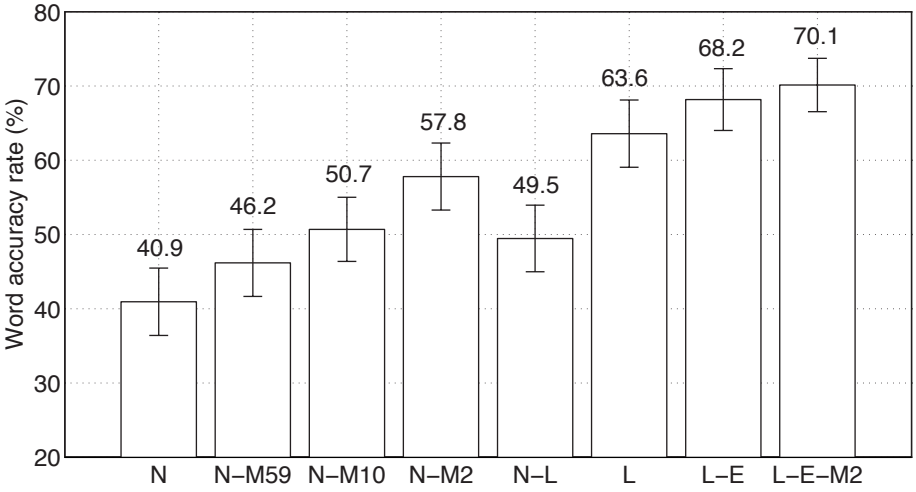


Figure 6.14: Word accuracy rates for speech-shaped noise.

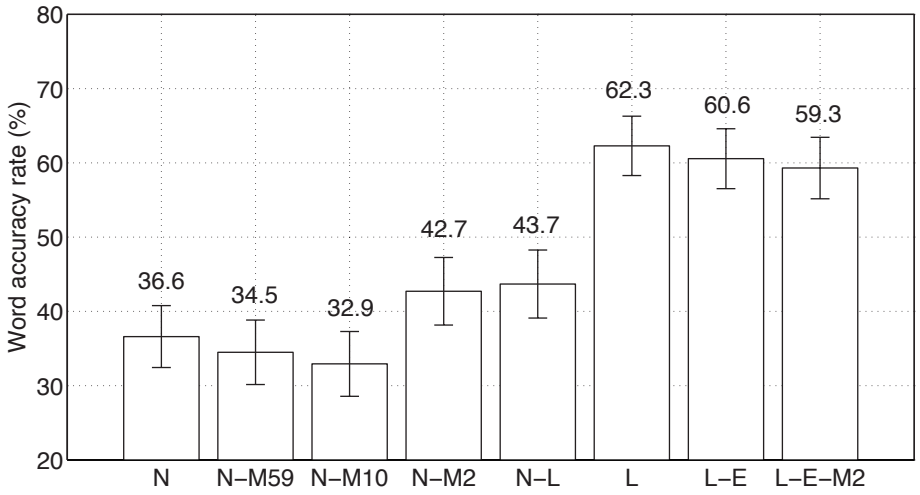


Figure 6.15: Word accuracy rates for competing speaker.

listening test. We used the first ten sets of the Harvard sentences (IEEE, 1969). The Harvard sentences are a group of 720 sentences organized in sets of 10, where each set is designed to be phonetically-balanced. The sentences are also more representative of everyday speech than the semantically unpredictable sentences used in other TTS intelligibility listening experiments (King and Karaiskos, 2010). Another one of the sets was used as a practice session done prior to the experiment. All words were considered when calculating the subjective word accuracy rate.

6.4.5 Results and discussion

Figs. 6.14 and 6.15 show the mean word accuracy rate (WAR) obtained by each voice when mixed with speech-shaped noise and a competing speaker respectively, along with 95 % confidence intervals. Fig. 6.14 shows that the modified voices N-M59, N-M10 and N-M2 achieve higher WAR than the unmodified voices N (40.9 %), and this is significantly higher for the N-M10 (50.7 %) and N-M2 (57.8 %). The N-M2 voice obtains a higher WAR than the N-L voice (49.5 %). The Lombard voices L (63.6 %), L-E (68.2 %) and L-E-M2 (70.1 %) performed better than the normal speech voices although we did not find a significant difference between N-M2 and L. The extrapolated voice L-E is more intelligible than voice L, a trend that is further enhanced by applying our modifications to it, as in voice L-E-M2. The results obtained for the competing speaker situation are displayed in Fig. 6.15 and show a slightly different trend. There is a drop in performance for N-M59 and N-M10 when compared to N (36.6 %), although this is not significant. The N-M2 (42.7 %) voice performs better than the unmodified counterpart N and obtains a similar WAR to N-L (43.6 %). All Lombard voices performed significantly better than the other voices, in particular the L voice (62.2 %). The other versions, L-E (60.5 %) and L-E-M2 (59.3 %), do not appear to increase intelligibility.

As predicted by our hypothesis that distortions were defeating potential gains in intelligibility in our previously-published experiments (Valentini-Botinhao et al., 2012a), the voices where we modify only the first few Mel cepstral coefficients achieved a better WAR, indicating that very fine frequency modifications cause distortions that cancel out any potential intelligibility gain they may offer. Compared to the N-L voice, for which the spectral parameters were obtained using Lombard speech, the modifications proposed here obtained a similar or higher intelligibility score. The intelligibility gains obtained by the full Lombard voice L over the N-L voice reflect the impact of changes in duration patterns, F_0 and the aperiodicity parameters that define the excitation signal, as pointed out in Table 6.2. We can see, then, that there is a lot to gain from modifying those parameters in addition to the spectral ones. The spectral modifications proposed here increased the gains obtained with the Lombard voice for speech-shaped noise, as we can see from the results for voice L-E-M2, which shows that there are still gains to be had over and above simply building voices on recorded Lombard speech.

For the competing talker, spectral changes seem to contribute less than for speech-shaped noise. For the competing talker, duration stretches as well as F_0 increases are

more important. This suggests that for non-stationary noise it is more effective to perform temporal energy re-allocation (e.g., taking advantage of quiet or silent regions in the noise signal) than it is to reallocate energy across different frequencies.

6.5 Conclusions

We have presented a method for increasing the intelligibility of HMM-generated synthetic speech in the presence of noise, based on the Glimpse Proportion measure. The method operates on the Mel cepstral coefficients generated by acoustic models which were trained only on natural read speech collected in quiet conditions, of the type normally used to build text-to-speech systems. The method updates the Mel cepstral coefficients iteratively via gradient descent such that the glimpse proportion increases, without changing the overall energy. We observed that sentences generated with such modified Mel cepstral coefficients have a boost in frequencies between 1-4 kHz and that this boost is highly dependent on the phonetic units: vowels and nasals are more enhanced than fricatives and stops. Results with a speech-shaped noise masker show that the modified voice is as intelligible as a synthetic voice trained with plain speech then adapted to Lombard speech. When mixed with a competing talker the gains are more modest for both the proposed method and for adaptation of Mel cepstral coefficients to Lombard speech.

In the next chapter, we provide a more extensive comparison with a wider variety of other intelligibility enhancement methods, and an investigation of method combination, particularly with methods that reallocate energy across time and change duration in a successful way.

Chapter 7

Evaluation of intelligibility enhancement methods

In this chapter, we present the result of three large scale listening experiments comparing the modification proposed in the previous chapter, referred now as GP, to other intelligibility enhancement methods applied to the same TTS baseline. Additionally, we evaluate a series of method combinations of GP with the following noise-independent methods: dynamic range compression, spectral shaper and adaptation to Lombard excitation and duration HMM models. The results presented as Evaluation I were obtained as part of a wider listening experiment, with entries for modifications applied to natural speech as well, described in Cooke et al. (2012). The results of Evaluation II were obtained from a listening test that we carried out using a similar design but comparing modifications applied only to TTS signals. We present the results of all voices involved in this evaluation. Evaluation III was part of Evaluation I follow-up experiment called the Hurricane Challenge (Cooke et al., 2013) with an even wider number of entries. For the first and third evaluations we present here only the results of our TTS entries and of the natural speech baselines. Our results in Evaluation I were partially published in (Valentini-Botinhao et al., 2012b), Evaluation II was published in (Valentini-Botinhao et al., 2013a) and our entries in Evaluation III were published in (Valentini-Botinhao et al., 2013d). Audio samples can be found at `https://wiki.inf.ed.ac.uk/CSTR/TtsHc` and `https://wiki.inf.ed.ac.uk/CSTR/HcExternal`.

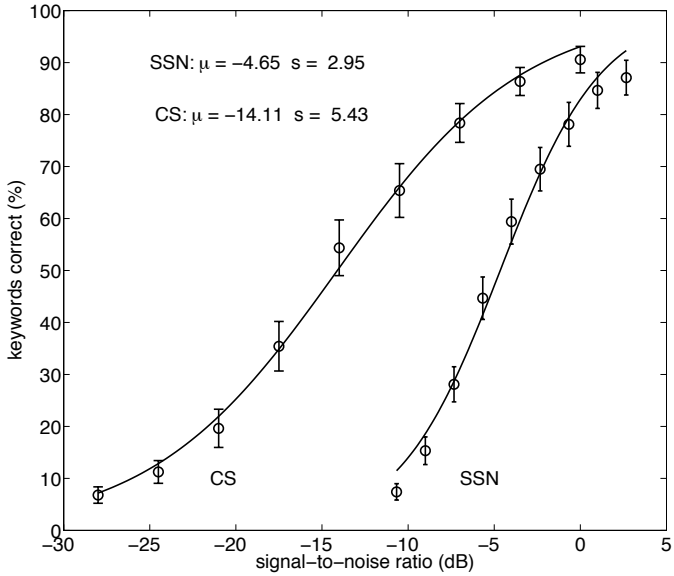


Figure 7.1: Psychometric curves obtained for natural speech in competing speaker and speech-shaped noise, figure extracted from Cooke et al. (2012). In this chapter, we refer to keywords correct as the word accuracy rate (WAR).

7.1 Stimuli material

The three evaluations we present here use the same TTS unmodified baseline as well as the same noise conditions, that is masker type and level. The TTS baseline used in the three evaluation was also used in the evaluation section of the previous chapter.

7.1.1 TTS

7.1.1.1 TTS baseline

The TTS baseline voice used here is voice N evaluated in the previous chapter. For details on how the TTS baseline voice was built see Section 6.4.1.

7.1.1.2 TTSGP

As we observed in the previous chapter, higher intelligibility gains were obtained when changing just the first two Mel cepstral coefficients, for this experiment we modify only these two coefficients. The voice that we call TTSGP here is the voice previously known as N-M2. More details on how this voice was created see Section 6.4.1.

7.1.2 Noise conditions

We mixed speech with two different maskers: speech-shaped noise and speech from a single competing female talker. The noises were mixed at preselected signal to noise ratios (SNRs) chosen to achieve approximately 25, 50 and 75% word accuracy rates (-9 dB, -4 dB, 1 dB for speech-shaped noise and -21 dB, -14 dB, -7 dB for competing talker). These SNR are referred here as 'Low SNR', 'Mid SNR', and 'High SNR'. The values were obtained from a separate listening test where natural speech intelligibility was evaluated in both noise maskers at nine different levels, obtaining the psychometric curves seen in Fig. 7.1. As these values were obtained using natural speech and not TTS we do not expect to have the same WAR values but do expect a similar spread in WAR.

For the GP modification, as in the previous chapter, no other energy normalization had to be performed to guarantee that the energy level of the sentence was not modified since modification does not modify the energy of the signal. For the samples generated with adaptation or dynamic range compression, the energy across a sentence had to be normalized to be equal to the plain unmodified synthetic voice generated sample.

7.1.3 Listening experiment

In all evaluations, the first 180 sentences of the Harvard corpus (IEEE, 1969) were used in a balanced arrangement, such that listeners never hear the same sentence more than once.

As in the other experiments presented so far, participants consisted mostly of undergraduate students from the University of Edinburgh, that is between 20 and 30 years old. All participants were British Native English speakers. Here we present the results across participants that had passed the audiological screening as described in Section 3.3.1. The participants heard stimuli presented over headphones in sound-isolated booths.

Results are presented as percent word accuracy rates (WAR) and equivalent intensity changes (EIC) in dB. The word accuracy rate was scored as the average across each listeners' individual scores for a particular voice in a particular noise condition, so the standard errors reflect listeners deviation rather than sentence material. Following the rules of the large scale evaluation (Cooke et al., 2012) the WAR scores were computed per sentence accounting only for content words (the words 'a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', 'for' were excluded from scoring) as oppose to counting all

words as was done in our previous evaluations. Although Cooke et al. (2012) refers to these scores as keyword correct rate we refer them here as WAR (%). We present here also the EIC, which is a relative measure of the performance of one voice compared to another. EIC is calculated by mapping the WAR scores that two voices obtain in a particular noise to the psychometric curve of that noise, as seen in Fig. 7.1 and calculating the effective change in dB (Cooke et al., 2012). Some results are reported with their corresponding standard Fisher's least significant differences, computed separately for each SNR level and masker type using ANOVAs with the single factor of modification type to allow for easier comparison across many modifications. Although not reported here, ANOVAs were computed with the single factor as the modification type and in all noise conditions significant differences were found across methods. We report in bar plots the standard error.

As done in the previous evaluations of this thesis, in this chapter we judge speech enhancement strategies with regard to their intelligibility benefit: naturalness or quality judgements were not obtained. This is because we focus here on finding the best technique in terms of intelligibility. Quality or naturalness could be used as a secondary criterion to compare techniques that provide similar intelligibility gains for instance. As was shown in the previous chapters, methods that degrade speech quality less can provide more intelligibility benefits. In that sense it is important to control the amount of perceived degradation. In this work this was done by limiting the amount of modification and providing to the listeners what we judged to be a reasonable quality.

7.2 Evaluation I: GP versus adaptation

We evaluated two natural voices plus three synthetic voices, whose acronyms are presented in Table 7.1. This evaluation was part of a large scale listening experiment described in Cooke et al. (2012) but in this section only we present the results of the synthetic speech entries as well as the results obtained for the natural and Lombard natural speech.

Using the same natural speech database described in our previous experiment we built three different voices for this evaluation: TTS, TTSGP and TTSLomb. The TTS and TTSGP voices here refer as said previously to the voices described as N and N-M2 in the previous chapter. The voice TTSLomb is built similar to the voice L but for this evaluation we limited the durations of voice L so that the maximum overall duration increase was no more than half a second per sentence. This has been done according

Voice	Modification	Adaptation to Lombard
Natural speech		
Normal	-	-
Lombard	-	-
Synthetic speech		
TTS	-	-
TTSLomb	-	all parameters
TTSGP	GP	-

Table 7.1: Evaluation I – voices.

	duration (s)	F_0 mean (Hz)	F_0 range (Hz)	spectral tilt (dB/oct.)	loudness (sone)
Natural speech					
Normal	2.06	107.1	34.60	-2.14	11.43
Lombard	2.32	136.8	46.74	-1.83	11.96
Synthetic speech					
TTS	1.95	104.5	22.45	-2.26	10.96
TTSGP				-1.90	12.43
TTSLomb	2.43	145.2	42.55	-1.71	12.06

Table 7.2: Evaluation I – Acoustic properties of the two natural voices: Normal and Lombard and the three synthetic voices: TTS, TTSGP and TTSLomb. These were calculated on a sentence level and averaged across the whole set of sentences.

to the rules of the extensive evaluation described in Cooke et al. (2012).

7.2.1 Acoustic analysis

In Table 7.2, we provide a sentence-level acoustic analysis of duration, fundamental frequency F_0 (mean and range), spectral tilt and loudness. To measure loudness we used the ISO-532B method (ISO 532, 1975), the F_0 range was calculated as the difference between the 80-th and 20-th percentiles and the spectral tilt was measured as the slope of the linear regression of the long term average spectrum on a one-third octave band scale as done previously. These values, presented in Table 7.2, are first calculated

per sentence and then averaged across the 180 sentences that were used in the listening test.

The natural Lombard sentences are on average 0.26 s longer than speech produced in quiet (a relative increase of 12%) and the synthetic Lombard TTSLomb sentences are 0.48 s longer (which corresponds to a relative increase of almost 25%). The mean fundamental frequency value F_0 is also higher for the Lombard voices, an increase of 27% and 39% for natural and synthetic speech respectively. The F_0 range also increases by 35% for natural and 90% for synthetic speech. Spectral tilt is found to be flatter: a relative change of 14% for natural and 24% for synthetic speech. The Lombard natural samples are on average 5% louder than normal speech ones and the Lombard synthetic voice TTSLomb is 11% louder than the normal synthetic voice TTS.

The voice built using the spectrum modification method TTSGP has the same duration and prosody as the TTS voice, but spectral tilt and loudness differ. The modified voice TTSGP presents a flatter spectral tilt when compared to the TTS voice (16% flatter) , though not to the same extent as the Lombard voice TTSLomb. The TTSGP is however slightly louder than the TTSLomb, a relative increase of 13% over the TTS voice.

The acoustic differences found here for the natural speech data are similiar to what has been reported in other studies of Lombard speech data described in Section 2.2: duration increases, F_0 mean and range increases, flatter spectral tilt and increase in loudness. A similar but stronger trend was observed for the synthetic voices.

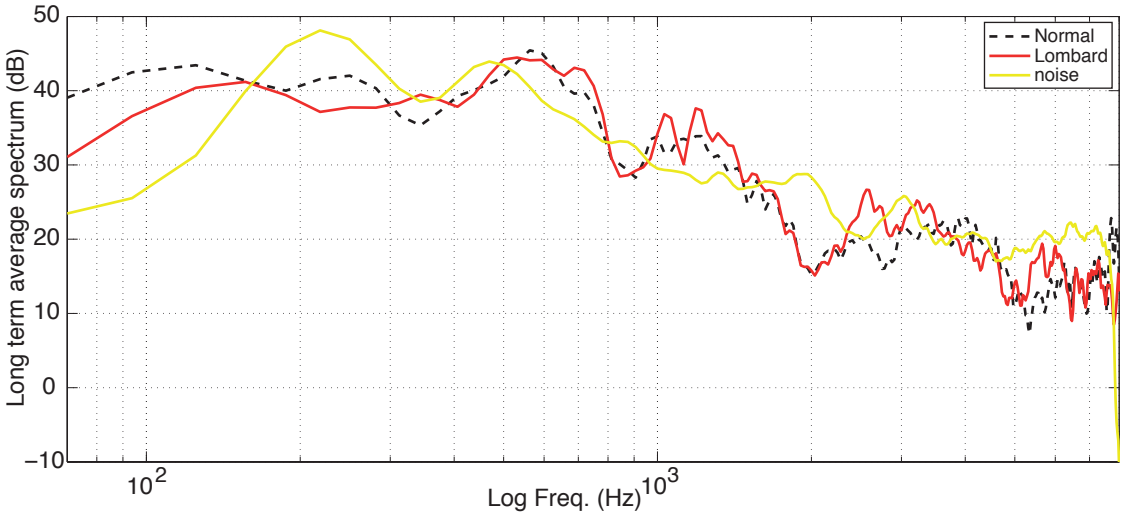


Figure 7.2: Long term average spectrum of one sentence of the natural voices.

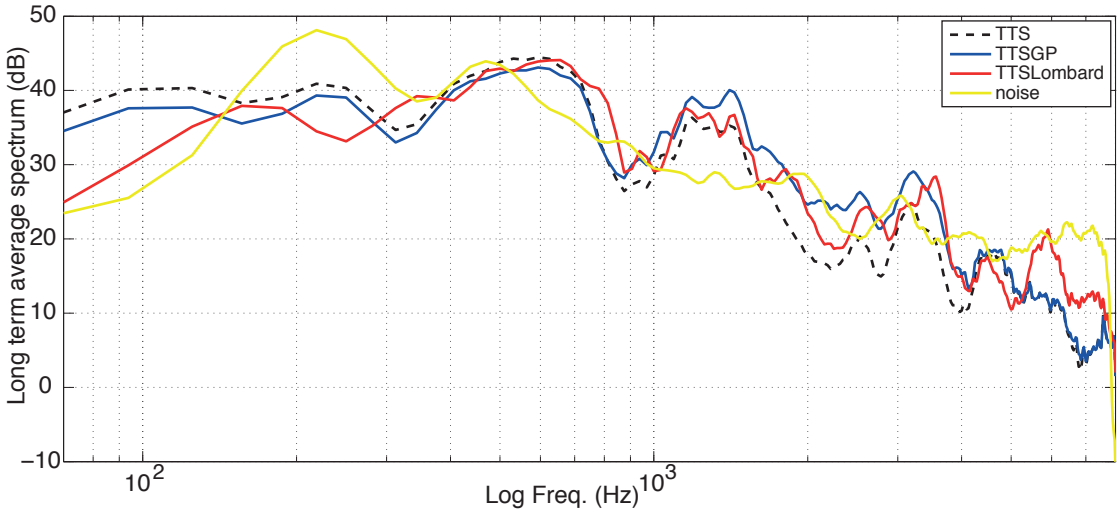


Figure 7.3: Long term average spectrum of one sentence from the TTS voices. Selected sentence was the same as in Fig. 7.2

Figs. 7.2 and 7.3 show the long term average spectrum (LTAS) of a sentence from the natural speech recordings and from the generated synthetic speech, respectively. In both figures we also display the LTAS of the noise that was used when creating the TTSGP voice: speech-shaped noise presented at -4 dB. We can see from Fig. 7.3 that, compared to the TTS curve, the curves for TTSLomb and TTSGP are attenuated at low frequencies, mostly below 1kHz and enhanced in the range above that. The TTSGP curve is mostly attenuated below 900 Hz and enhanced in the region between 900-4000 Hz. The TTSLomb voice curve is less pronounced in this region but shows a boost in the region above 5 kHz. We can also see this effect in the natural Lombard

speech curve displayed in the Fig. 7.2. The Lombard voices, both natural and synthetic, also present a shift in fundamental frequency and formants.

7.2.2 Listening experiment

The listening test involved 154 native English speakers, 15 of which did not pass the audiological screening so only the results of 139 participants were considered (Cooke et al., 2012). Each participant listen to 4 different sentences of each noise/SNR/voice combination.

7.2.3 Results and discussion

Fig. 7.4 shows the word accuracy rates (WAR) of speech mixed with speech-shaped noise (SSN) and competing speaker (CS) for each SNR tested.

An obvious first comparison to draw is the difference in performance gain when using natural and synthetic Lombard speech. Averaged across the three different SNRs the gains in intelligibility obtained by the Lombard synthetic voice TTSLomb over the normal synthetic voice TTS are larger (47% for SSN and 42% for CS) than the gains obtained by the Lombard natural speech over the normal natural speech (17% for SSN and 13% for CS). The effects are most pronounced for the lower SNRs cases for speech-shaped noise and for the middle SNR case for the competing talker condition.

The noise played when recording the Lombard dataset used in this evaluation was different to the ones used in the listening test. We can thus infer that Lombard speech can still be more intelligible than speech produced in quiet even in a mismatched scenario. This could indicate that certain modifications can provide improvements independent of the noise.

Most importantly we see that the GP-based Mel cepstral modification (TTSGP) can provide intelligibility gains over the non-Lombard synthetic voice (TTS). The word accuracy rates obtained by the TTSGP voice are comparable to those obtained with the TTSLomb voice for speech-shaped noise even though no modification was made to duration or to the excitation signal. Averaged across SNRs, the relative gains obtained over the TTS voice were 44% for SSN and 5% for CS. For the competing talker only moderate improvements were obtained by TTSGP over TTS, suggesting a greater importance of prosody and duration in this scenario.

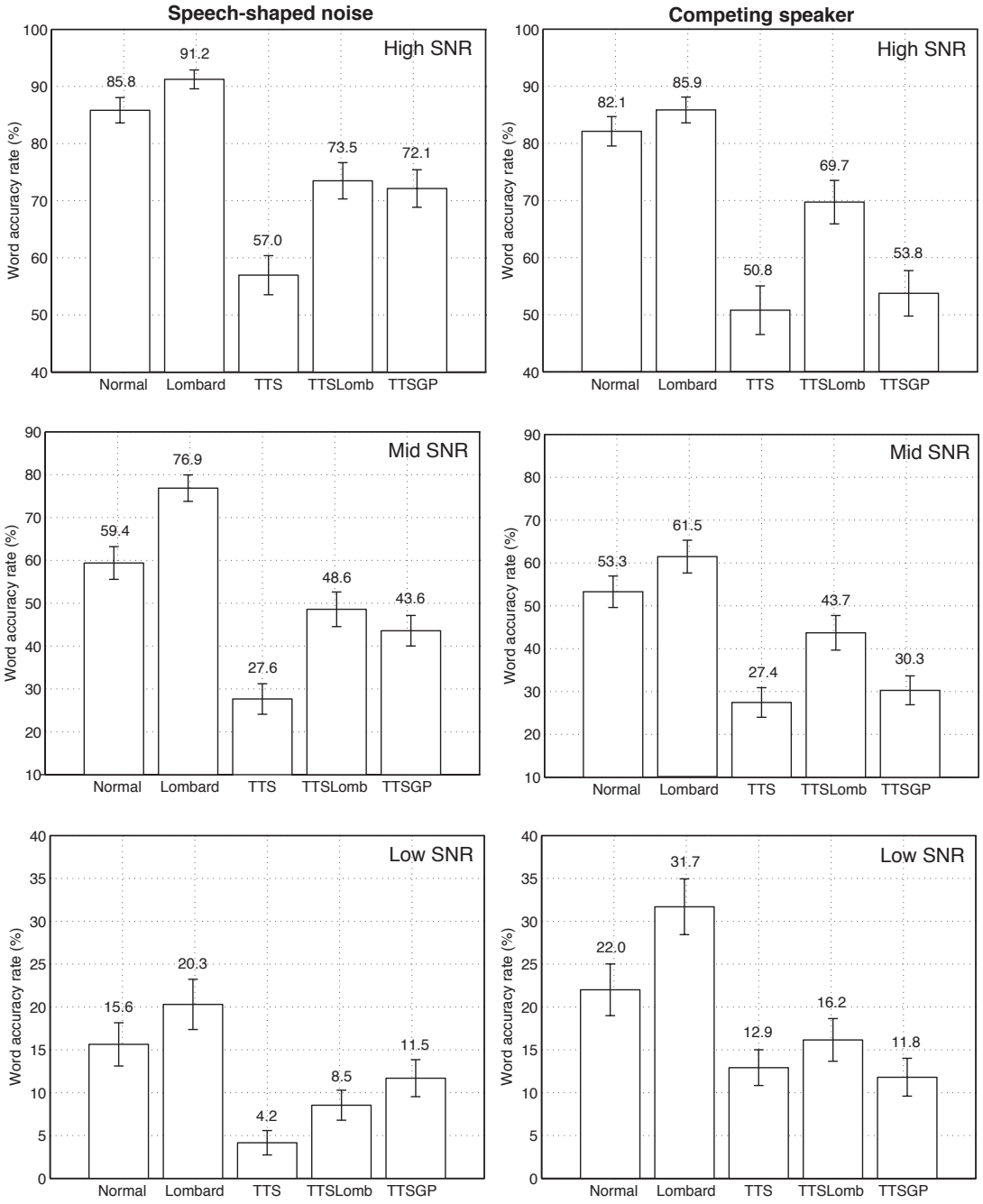


Figure 7.4: Evaluation I – Word accuracy rates for natural voices (Normal, Lombard) and synthetic voices (TTS, TTSLomb, TTSGP) mixed with speech-shaped noise (left) and competing speaker (right) at the conditions high SNR (top), mid SNR (middle) and low SNR (bottom).

The TTS voices obtained lower WAR when compared to natural voices. On average across different noises and SNRs the TTS voice WAR is 23% lower than natural speech and TTSLomb WAR is 18% lower than the Lombard voice.

The method employed in Cooke et al. (2012) uses a psychometric function which means we are able to express the change in intelligibility in terms of "equivalent intensity change" (EIC) relative to normal natural speech, which is an intuitively appealing way of presenting the results on a dB scale. This is shown in Fig. 7.5 for SSN and CS. We can see the effective loss (in dB) of using synthetic speech compared to natural speech (average across SNR: TTS -4.3 dB for SSN and -5.9 dB for CS) and how this loss can be substantially mitigated by modifying the synthetic voice spectral envelope using our proposed method (TTSGP -1.8 dB for SSN and -5.6 dB for CS) or by adapting the models to Lombard speech from the same speaker (TTSLomb -1.9 dB for SSN and -2.7 dB for CS).

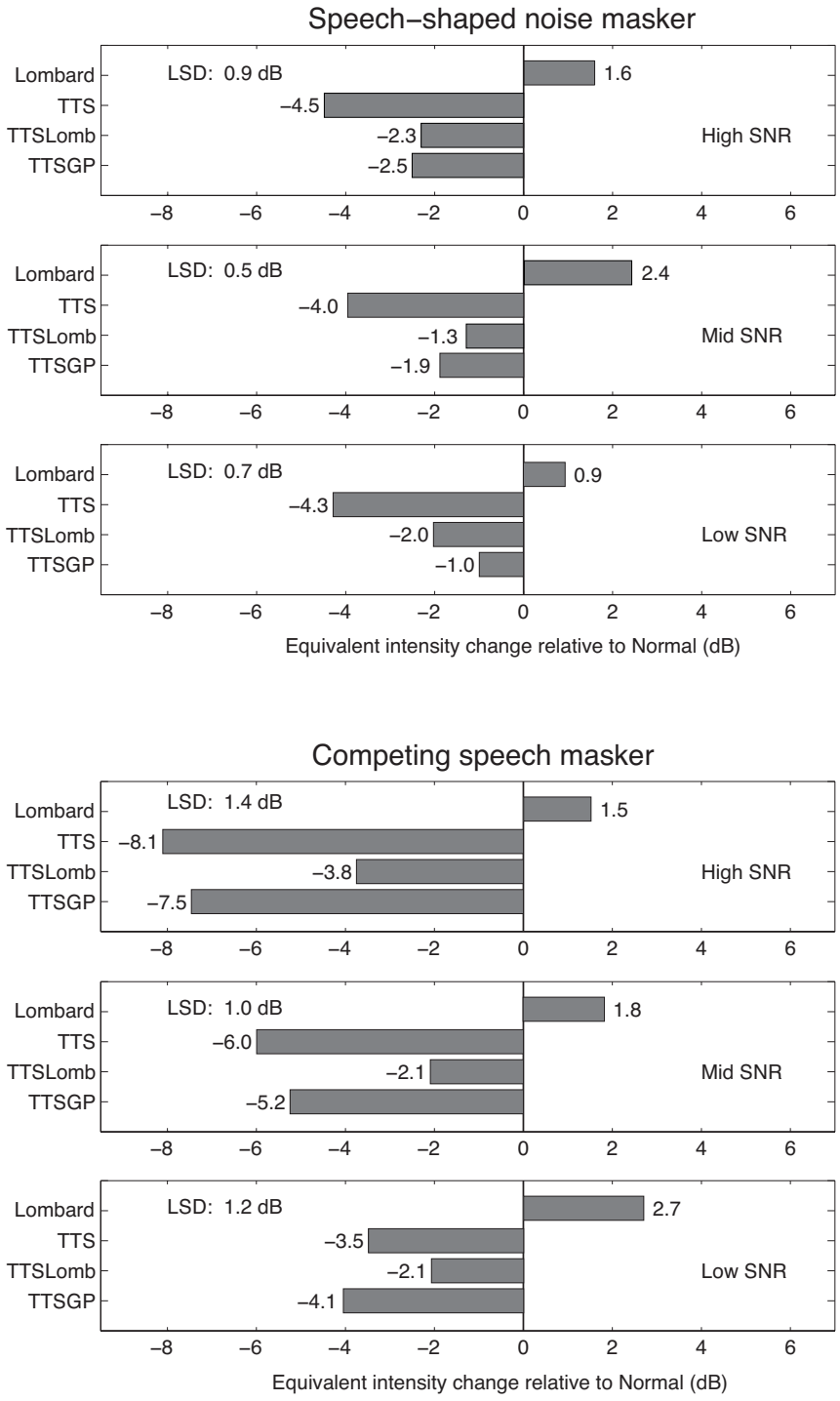


Figure 7.5: Evaluation I – Equivalent intensity change relative to natural speech, for speech-shaped noise (top) and competing speaker (bottom). LSD indicates Fisher's least significant difference converted to dB via the psychometric function for this masker. Figures adapted from (Cooke et al., 2012).

7.3 Evaluation II: Noise-dependent and -independent methods

While TTS voices can be as intelligible as natural speech in clean conditions, intelligibility drops quite rapidly in adverse conditions (King and Karaiskos, 2010), motivating the use of intelligibility enhancement methods and potentially requiring knowledge of the noise masker. However, noise-dependent methods, either for natural or TTS voices, have only recently been proposed and it remains relatively unknown to what extent exploiting spectro-temporal characteristics of the masker is useful. To evaluate a range of enhancement algorithms, both noise-dependent and independent, Cooke et al. (2012) describes a large scale listening experiment with 5 methods for natural speech and 2 for TTS evaluated under the same conditions and which results were presented in the previous section. In this evaluation, it was observed that noise-independent spectral shaping with Dynamic Range Compression (SSDRC) (Zorilă et al., 2012) provided the best results of the modifications on natural speech while the method proposed in Sauert and Vary (2011), although noise-dependent, did not perform as well. We observed in the previous section that a noise-dependent approach applied to a TTS voice can produce a voice that is as intelligible as a Lombard-adapted voice in some stationary noise conditions, but still not as intelligible as natural speech. A significantly large intelligibility gap was also confirmed between TTS and the natural voice in almost all noise conditions.

In this section, we investigate whether intelligibility enhancement methods originally proposed for natural speech can also improve intelligibility of a TTS voice and help bridge this gap. Furthermore, we seek to discover whether it is possible to improve a noise-independent method (Zorilă et al., 2012) and the noise-dependent method described in the previous chapter by combining them, effectively offering insight on the extent to which noise dependency is required in terms of achieving significant intelligibility gains.

7.3.1 Methods

We evaluate one natural voice and a total of seven TTS modified voices, as shown in Table 7.3: two noise-independent methods (SS-DRC (Zorilă et al., 2012) and SSE-DRC), two noise-dependent methods (TTSGP and OptSII (Sauert and Vary, 2011)) and two method combinations (TTSGP-DRC and TTSGP-SS-DRC). The TTSGP method

Voice	Modification	ND
Natural speech		
Normal	-	-
Synthetic speech		
TTS	-	no
TTS-SS-DRC	spectral shaping (SS) followed by dynamic range compression (DRC) (Zorilă et al., 2012)	no
TTS-SSE-DRC	extended version of SS (SSE) followed by DRC	no
TTS-OptSII	SII optimisation (Sauert and Vary, 2011)	yes
TTSGP	GP	yes
TTSGP-DRC	GP followed by DRC	yes
TTSGP-SS-DRC	GP followed by SS-DRC	yes

Table 7.3: Evaluation II – voices, ND stands for noise dependency.

is applied directly to the generated spectral parameters, all other methods work as a post processing of the waveform generated by the TTS model (represented by the addition of the acronym TTS-). The following describes each of the methods in more detail.

SS-DRC (Zorilă et al., 2012) performs spectral shaping (SS) followed by dynamic range compression (DRC). Spectral shaping consists of two cascaded subsystems which are adaptive to the probability of voicing: (i) an adaptive sharpening where the formant information is enhanced, and (ii) an adaptive pre-emphasis filter. A third fixed spectral shaping is used to prevent attenuation of high frequencies in the speech signal during the signal reproduction. The output of the spectral shaping system is then input to the DRC, inspired by compression strategies used in sound recording and reproduction, audio broadcasting as well in amplification techniques in hearing aids (Blesser, 1969).

The extended spectral shaping (SSE) is carried out on all voiced frames and consists of three components: (i) a fixed filter to increase the spectral energy gain in certain frequency bands, (ii) peak enhancement via cepstral liftering and (iii) slight formant shifting via frequency warping. First, the fixed filter is bi-modal, with the most gain (12 dB) between 1-4 kHz, mimicking the spectral gains observed in Lombard speech (Godoy and Stylianou, 2012), and the secondary mode has approximately half of the maximal gain and is concentrated between 5.5-7.5 kHz. Second, the peak-enhancement follows the peak-weighted cepstral lifter (α=0.85) presented in Kim and

Lee (2000) for enhancement in speech recognition. Third, the frequency warping shifts the first and second formants moderately (less than 100 Hz) on average upwards in frequency. The frequency warping function is constant and derived from observations on the expanded vowel space of two speakers in a separate clear speech corpus involving the Harvard sentences.

In the OptSII method (Sauert and Vary, 2011, 2012) the audio power of the speech signal is spectrally reallocated with respect to the speech intelligibility index (SII) (ANSI, 1997). A recursive closed-form optimisation scheme calculates, for each time frame, the spectral weights in 21 Bark-scaled subbands which maximise the SII, given the current disturbance spectrum levels, with the additional constraint of an unchanged short-term audio power of the speech signal. Opposed to Sauert and Vary (2011) and the OptSII style used in Cooke et al. (2012), in this evaluation a moving average noise estimator is used, which is also able to track CS noise.

7.3.2 Acoustic analysis

As all methods modify the speech spectrum, while maintaining prosody and duration, we present here acoustic analysis based only in terms of spectral gains. Similar to the phone level acoustic analysis presented in the previous chapter, we calculated these gains at the phone level and grouped the results in phone classes: vowels, nasals, approximants, stops, fricatives and affricates as seen in Figs. 7.6 and 7.7. To obtain these gains, we calculated the phone periodogram by extracting a 512 point discrete Fourier transform calculated using a 20 ms hamming window at every 10 ms and averaged across the time frames within the phone boundaries. The gain is then the difference of the phone periodogram in dB for a certain method and the periodogram for the unmodified TTS speech. For the noise-dependent methods, this was calculated for speech-shaped noise (SSN) in the mid SNR condition: results will differ for other noise types and levels.

For the majority of the methods, the average spectral gains can be interpreted as a sort of correction filter that re-allocates spectral energy and remains largely constant across phones for stationary maskers like the SSN. For TTS-SS-DRC, the gain curve shape is determined primarily by the SS fixed filter (seen for all phones), but the effective scale of the gains is affected by the DRC. As we can see in Figs. 7.6 and 7.7, the SS fixed-filter has a very wide-flat gain between 1-4 kHz and a gradual rolloff with increasing frequency. The gain curves in TTSGP, TTS-SSE-DRC and TTS-OptSII, on

the other hand, are generally bi-modal. Note that the shape of the fixed-filters or gain curves is most apparent with the voiced phones, particularly with vowels. As observed in Sauert and Vary (2012) at low SNR, OptSII shows a bandpass characteristic, at mid SNR the general spectral shape of the speech signal tends to follow the shape of the noise, and at higher SNR the spectral gains are quite low.

When comparing TTSGP and TTSGP-DRC, we can clearly see the effect DRC has: gain reduction especially on vowels and increased gain on stop and fricatives, while also determining an upward-sloping linear-like gain curve shape on these last phones. That is, DRC is re-allocating energy of frames in such a way as to increase loudness of the unvoiced parts of speech.

Looking at the gain curves for the TTSGP-SS-DRC method, we can see that the fixed-filter shape of SS dominates, but the GP gain curve is apparent in the roundness of the first mode in the voices (first three categories). More importantly, the scale of the gain is compounded by combining the GP-SS as seen from the gain obtained on the voiced segments.

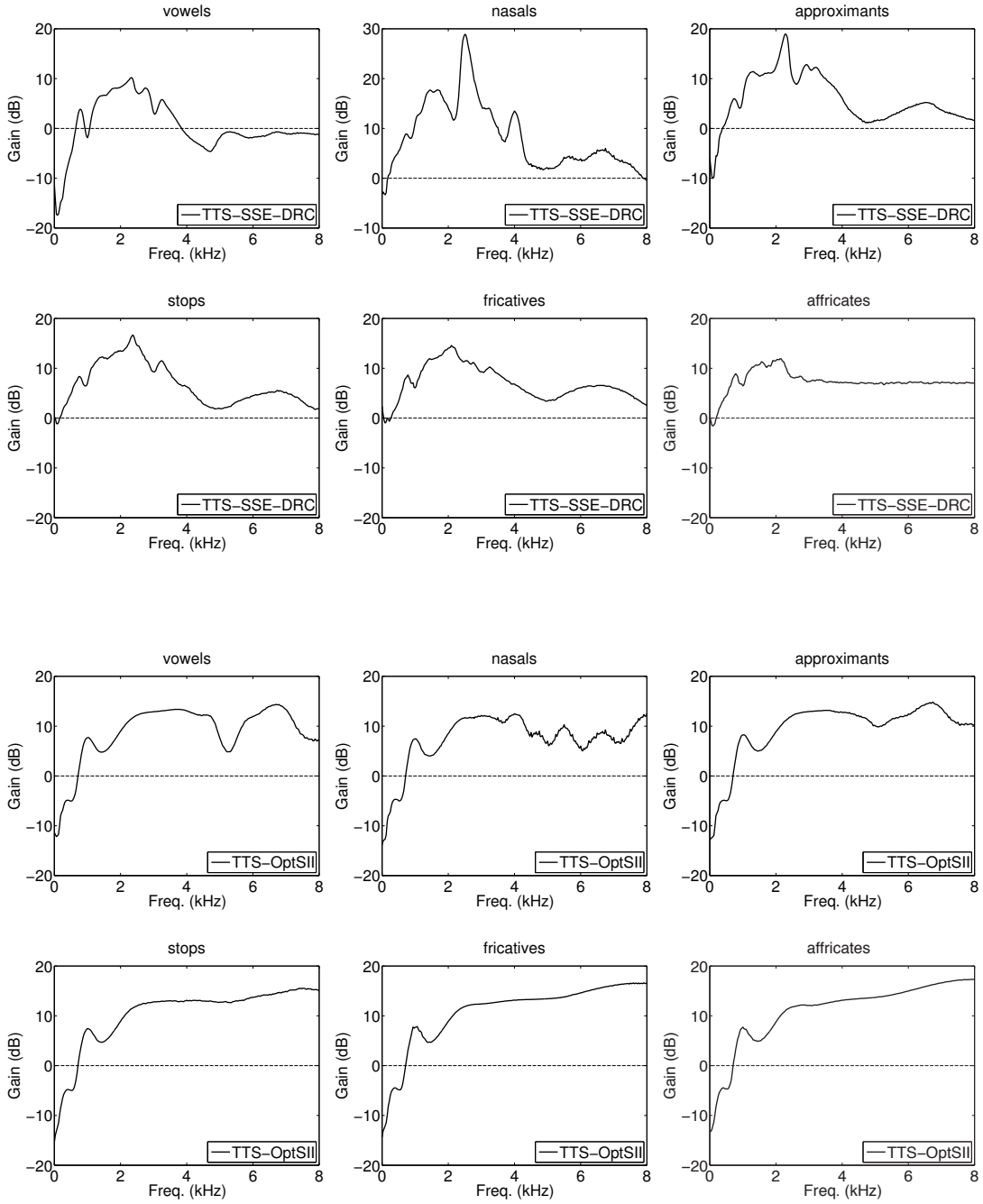


Figure 7.6: Evaluation II – Spectral gains (dB) exhibited by each voice compared to the unmodified baseline TTS. Voice TTS-SSE-DRC (top) and TTS-OptSII (bottom).

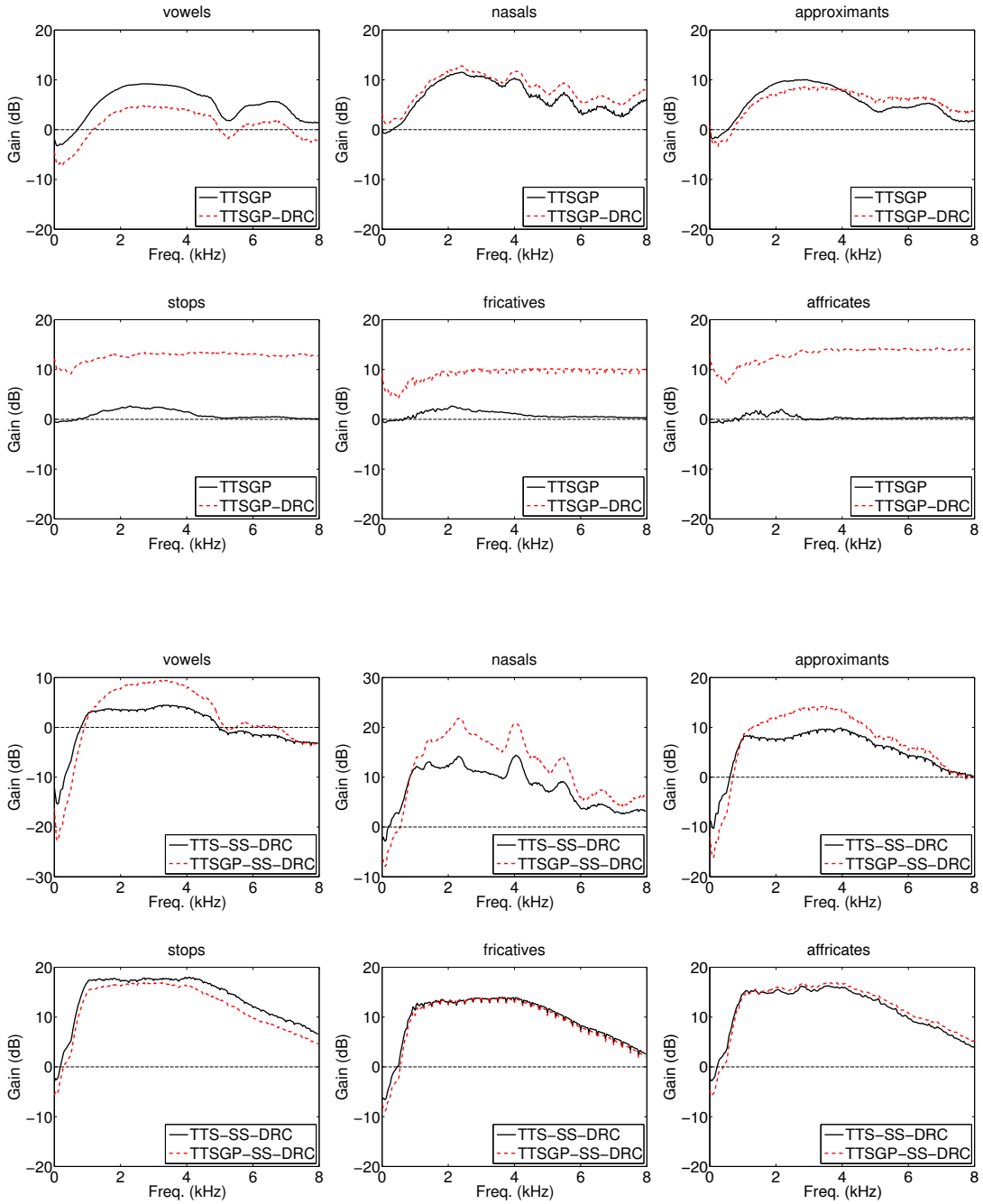


Figure 7.7: Evaluation II – Spectral gains (dB) exhibited by each voice compared to the unmodified baseline TTS. Voices TTSGP & TTSGP-DRC (top) and TTS-SS-DRC & TTSGP-SS-DRC (bottom).

7.3.3 Listening experiment

In order to obtain listening scores for word accuracy, we performed a listening test with 88 native English speakers. Groups of 4 participants heard 15 different sentences for each listening condition (noise/SNR/voice).

7.3.4 Results and discussion

We evaluated the seven different TTS styles displayed in Table 7.3 as well as natural speech in a large listening experiment. We present intelligibility scores in terms of word accuracy rates (WAR) in % and equivalent intensity change (EIC) in dB.

Fig. 7.8 shows the WAR calculated across all sentences for each voice in each listening condition. The dashed line corresponds to the WAR obtained for the natural speech in that condition. The results are organized by noise type and level. Fig. 7.9 presents the EIC in dB relative to natural speech.

Overall, we can see that the most effective method is the TTS-SS-DRC. It seems that the width and the gain of the primary mode is very important as TTSGP, TTS-SSE-DRC and TTS-OptSII suffer slightly as a result of too narrow or curved gains indicating that, for example, the gains around 1 kHz should be higher. The secondary mode does not seem to benefit intelligibility as much.

7.3.4.1 DRC effect

All methods except the TTSGP perform some sort of high frequency boosting which enhances voiced segments. This significantly aids intelligibility in the SSN and CS conditions as these noises have stronger low frequency components. We can clearly see this intelligibility gain by comparing the results of TTSGP-DRC and TTSGP: DRC improves TTSGP performance in all noisy conditions, particularly for the SSN Mid SNR condition. TTSGP-DRC and TTS-OptSII obtained similar performance: in SSN TTSGP-DRC performs better in the mid and high conditions and no significant differences appeared in CS. At lower SNRs, larger gain at higher frequencies (observed for TTS-SS-DRC and TTS-OptSII) seems to be more beneficial most likely due to the masker.

7.3.4.2 Combination effect

Applying SS-DRC to a TTSGP style voice did not improve the intelligibility as we see that TTSGP-SS-DRC either obtained worse or similar WARs compared to TTS-SS-DRC. The compounded gain of SS-GP seen in the acoustic analysis is most beneficial at Low SNR (specifically for SSN). Otherwise, it seems excessive and TTSGP-DRC or TTS-SS-DRC are sufficient.

A few methods were as intelligible as natural speech in SSN Mid SNR and Low SNR conditions. TTS-OptSII, TTS-SS-DRC and TTSGP-SS-DRC were significantly more intelligible than natural speech in SSN Low SNR. For the CS case TTS-SS-DRC was as intelligible as natural speech at Low SNR. Natural speech in CS for all SNRs was significantly more intelligible than the TTS styles. The differences among the methods is attenuated in CS, that is the gains obtained by the noise-independent method TTS-SS-DRC were attenuated.

These results can be converted to equivalent intensity changes (EIC) relative to normal natural speech on a dB scale as proposed by Cooke et al. (2012). Calculating this with respect to natural speech, see Fig. 7.9, we found that TTS-SS-DRC was 2.0 dB more intelligible than natural speech in SSN Low SNR; this gain is lower for the Mid SNR condition: 0.7 dB. A higher gain was obtained by TTSGP-SS-DRC: 2.25 dB in SSN Low SNR. The gap between modified TTS and natural speech is larger for CS; for Mid and High SNR conditions. Most methods were at least 4 dB less intelligible than natural speech while for the Low SNR condition TTS-SS-DRC decreased the gap to -0.7 dB.

We saw that a few methods made the TTS voice even more intelligible than the natural one. Although noise-dependent methods did not improve gains, the intelligibility differences found in distinct noises motivates such dependency.

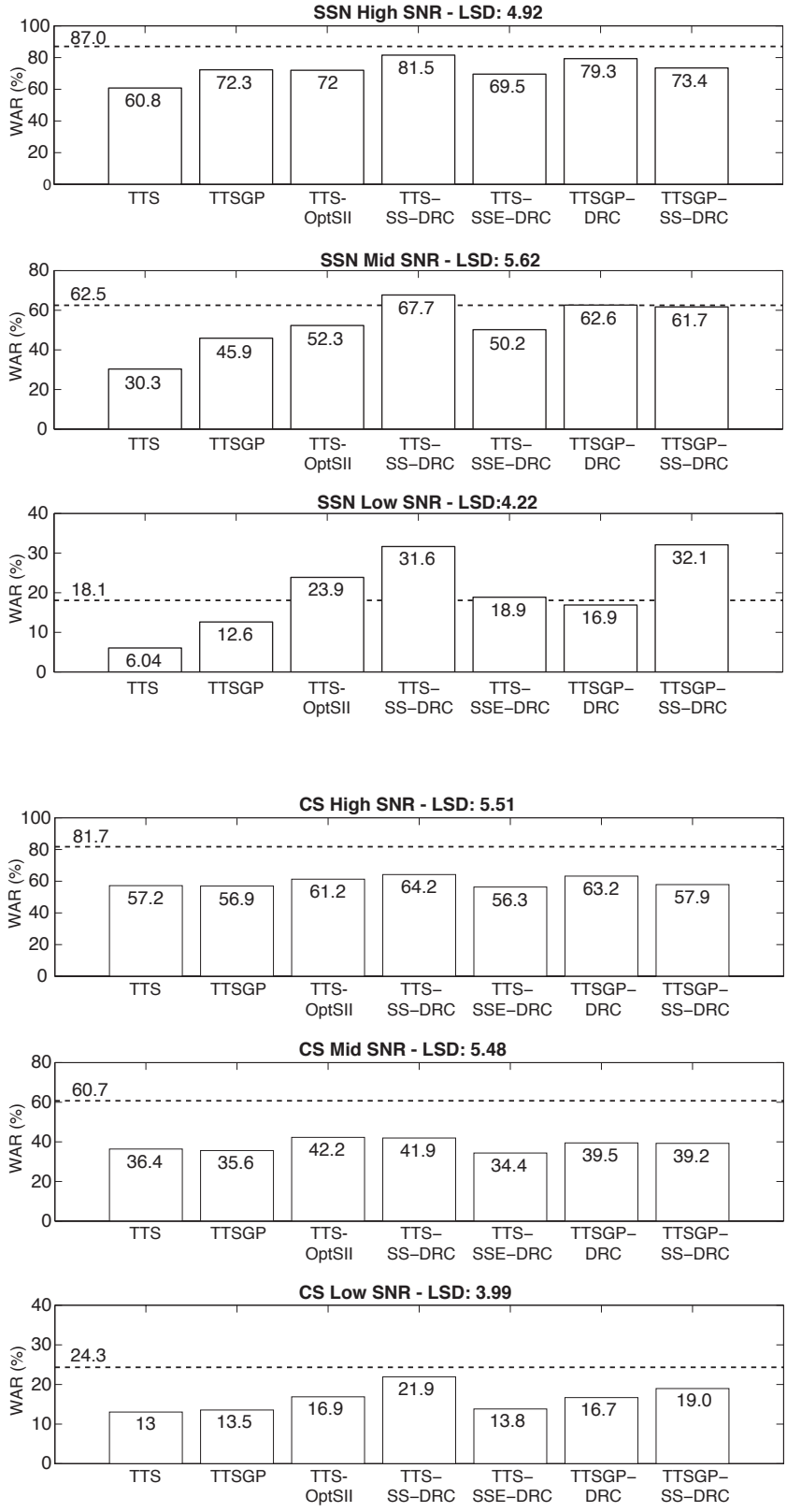


Figure 7.8: Evaluation II – Word accuracy rate (WAR) obtained in the listening evaluation for speech-shaped noise (top) and competing speaker (bottom). The dashed line corresponds to the WAR obtained for natural speech in that condition. LSD is Fisher's least significant difference.

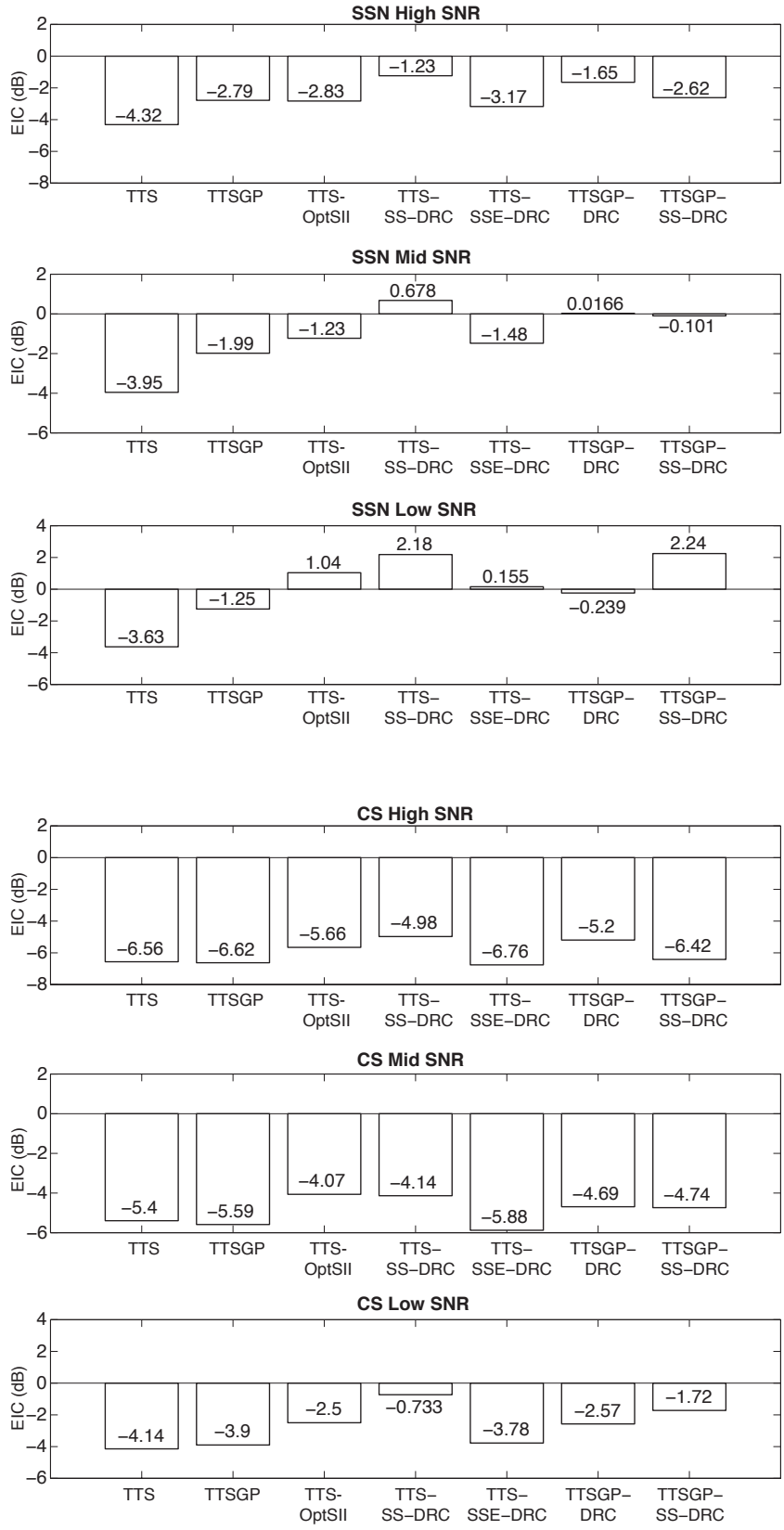


Figure 7.9: Evaluation II – Equivalent intensity change (EIC) relative to natural speech, for speech-shaped noise (top) and competing speaker (bottom).

7.4 Evaluation III: Adaptation and noise dependency

In the previous chapter, we found that is possible to obtain larger intelligibility gains by performing spectral modifications, than by adapting a plain speech TTS spectral model to Lombard data. Moreover, we found in the previous section of the current chapter that dynamic range compression can further boost this gain. Although we obtained substantial gains in speech-shaped noise, our results in the case of a competing speaker were not as good. To improve this performance, we propose to incorporate duration and excitation changes from Lombard speech, by combining three different modification strategies: spectral changes based on the glimpse proportion measure (GP), dynamic range compression (DRC) and adaptation to Lombard duration and excitation.

Although observed in natural Lombard speech (Junqua, 1993; Lu and Cooke, 2008; Hazan and Baker, 2011), reproducing changes in duration and fundamental frequency (F_0) does not necessarily generate significant intelligibility gains (Lu and Cooke, 2009b; Villegas et al., 2012). In Chapter 4, we manipulated the duration and F_0 of a TTS voice and no significant increases in intelligibility were observed when increasing F_0 in the four noise types tested (car, high frequency, speech-shaped and cafeteria). Slowing the speaking rate resulted in a few significant gains in the speech-shaped noise and cafeteria masker. In Chapter 6, however, we saw that quite a significant gain came from using Lombard-adapted fundamental frequency and duration in the competing speaker scenario, even though the noise used for inducing such changes was not matched to the competing speaker masker. A combined solution for improving results in competing speaker noise while maintaining the gains already achieved in speech-shaped noise, is to use Lombard-derived excitation and duration changes (noise-dependent but not matched) through voice adaptation (Yamagishi et al., 2009), apply the GP-based spectral shaper and follow this by DRC. We refer to this combination of strategies as the TTSLGP-DRC voice.

As previously stated, the results shown here are derived from a wider evaluation called the Hurricane Challenge, which compared other types of speech intelligibility enhancement methods applied to natural and TTS (Cooke et al., 2013). Results from all entries are reported separately in Appendix C.

Voice	Modification	Adaptation to Lombard
Natural speech		
Normal	-	-
Synthetic speech		
TTS	-	-
TTSLGP-DRC	GP followed by DRC	excitation and duration

Table 7.4: Evaluation III – voices.

7.4.1 Methods

Voice TTSLGP-DRC was based on voice TTS but the models for duration and excitation were further adapted. The two first Mel cepstral coefficients were modified using the method proposed in Chapter 6 and we applied a dynamic range compressor (DRC) (Zorilă et al., 2012) to the synthesized waveform.

According to the rules of the Hurricane Challenge, each sentence can not be longer than its corresponding noise file, as provided by the challenge, which is around one second longer than the corresponding natural speech signal. To keep within this rule, we had to restrict the duration of the generated sentences, because otherwise they would have been on average 0.69 s. longer than the natural speech, with a significant number of sentences more than one second longer than natural speech. We decided to restrict the duration of each generated sentence to be no more than 300 ms longer than the corresponding natural speech, to allow 300 ms leading / 200 ms lagging noise signal before/after the stimuli presented to the listeners. To achieve this, we forced the overall duration of the sentence to be within this rule (when necessary) (Yoshimura et al., 1998). Because changing the overall duration of the sentence does not actually guarantee a sufficiently reduced duration (due to rounding errors mentioned in Section 3.2.4), we iteratively decrease the duration (in steps of 100 ms) until it was within the required limits. In the final stimuli, the average duration difference (compared to natural speech) was 0.45 s with only once sentence above the 0.5 s limit (0.53 s).

7.4.2 Acoustic analysis

To give more insights into the results, we provide in Table 7.5 a sentence-level acoustic analysis of duration, fundamental frequency F_0 (mean and range), spectral tilt and loudness (measured using the ISO procedure), calculated in the same manner as Table 7.2 described in Section 7.2.1.

	duration (s)	mean/range F_0 (Hz)	spectral tilt (dB/oct.)	loudness (sone)
Natural speech				
plain	2.06	107.1 / 34.60	-2.14	11.43
Lombard	2.32	136.8 / 46.74	-1.83	11.96
Synthetic speech				
TTS	1.95	104.5 / 22.45	-2.26	10.96
TTSGP			-1.90	12.43
TTSGP-DRC			-1.45	13.37
TTSLGP-DRC	2.49	145.2 / 42.55	-1.46	13.12
TTSLomb	2.43		-1.71	12.06

Table 7.5: Evaluation I, II and IIII – Acoustic properties at sentence level averaged across the dataset of the two natural voices: Normal and Lombard and the five synthetic voices: TTS, TTSGP (Evaluation I), TTSGP-DRC (Evaluation II), TTSLGP-DRC (Evaluation III) and TTSLomb (Evaluation I).

Fig. 7.10 shows the long term average spectrum calculated per sentence and averaged across sentences, for some of the TTS voices.

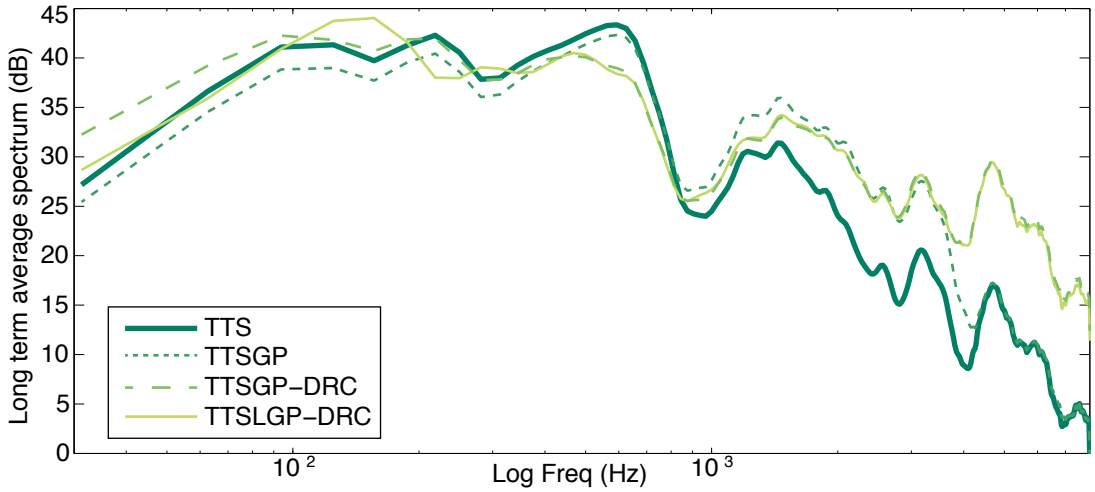


Figure 7.10: Long term average spectrum calcualted at a sentence level and averaged across the dataset.

We see a tendency for GP and DRC to increase the loudness of speech and flatten the spectral tilt, even more so than for the Lombard natural voice. Changes in the spec-

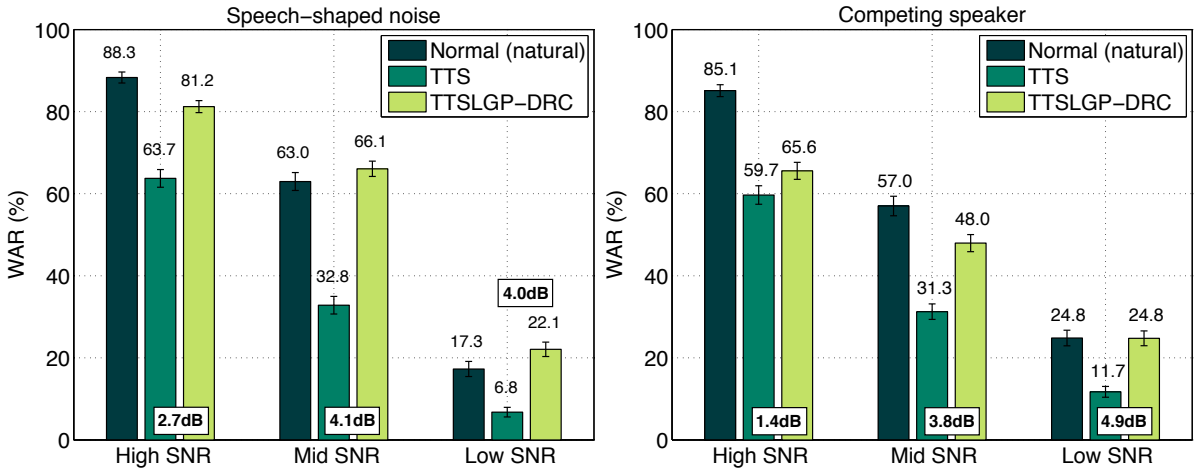


Figure 7.11: Evaluation III – WAR and EIC relative to TTS baseline (values inside boxes) for speech-shaped noise (left) and competing speaker (right).

tral tilt observed when DRC is applied are a consequence of the temporal modifications this method performs. DRC boosts regions of the speech waveform that are of a lower level and these regions correspond to the higher frequency components of the speech spectrum, i.e. a flatter spectrum. We see the boosting effect that GP has around the formant frequency range, and the boosting that DRC gives to higher frequencies. F_0 and its range (within a sentence) is increased in voices built with Lombard excitation.

7.4.3 Listening experiment

In total 175, native English speakers participated in the listening test. Each participant transcribed 9 different stimuli per voice.

7.4.4 Results and discussion

Fig. 7.11 shows word accuracy rates (WAR) and standard errors for the synthetic voices TTS and TTSLGP-DRC and the natural plain speech entry, mixed with speech-shaped noise (left) and competing speaker (right). In all noise conditions, the gap between natural and TTS is smaller with the TTSLGP-DRC voice, particularly for the lower SNR cases in both noise types.

Fig. 7.11 also presents EIC gain values inside boxes. Relative to TTS, the proposed voice achieved gains of 2.7, 4.1 and 4.0 dB in the SSN condition, going from highest to lowest SNR. When mixed with a CS the gains are 1.4, 3.8 and 4.9 dB. If we compare

the Mid and Low SNR gains across noises on the dB scale they appear more similar than on the WAR scale.

7.5 Comparing across different listening tests

To be able to understand the contribution of each component (GP, DRC and Lombard excitation & duration), we compared the changes relative to natural speech for each of the TTS voices described in Table 7.5. We show these results expressed in WAR and EIC (values inside boxes) in Fig. 7.12.

As all the listening tests had the same set-up (sentence material, noise types and SNRs, stimuli presentation) and same scoring rules; we can compare results across them by calculating the gains that each modification obtained relative to the results that the natural speech entry obtained in that particular test. Similar to the scoring methodology used thus far, we present the WAR change averaged across the gains obtained by each voice for each listener, and again this means that the standard error measures the variability across listeners. The number of points that define the standard error is defined by the number of participants: 139 in Evaluation I, 88 in Evaluation II and 175 in Evaluation III. As the TTS entry was present in all three experiments, we show the WAR change for that system averaged across all participants (402 points).

When comparing the voices TTS, TTSGP, TTSGP-DRC and TTSLGP-DRC we can see the gain that each component adds. This addition depends on the noise type and SNR, meaning that some components are more important in one condition than another. In speech-shaped noise, as shown in the top part of Fig. 7.12, GP and DRC contribute most. Duration and excitation changes start contributing only at quite low SNRs. The picture is different for the competing speaker condition in the lower part of Fig. 7.12, where GP and DRC gains are quite modest (apart from the significant gain observed in the highest SNR condition for DRC – where the masker is more an energetic masker than an informational one) and the Lombard-based changes contribute most for the Mid and Low SNR conditions, where 'filling the gaps' in time/frequency is more beneficial than being louder (as seen in Table 7.5).

Comparing TTSLGP-DRC and TTSLomb we can see the additional gain that GP and DRC provide over adapting spectral parameters as well as duration and excitation parameters, particularly for the mid and low SNR conditions of SSN and for the low SNR for competing speaker.

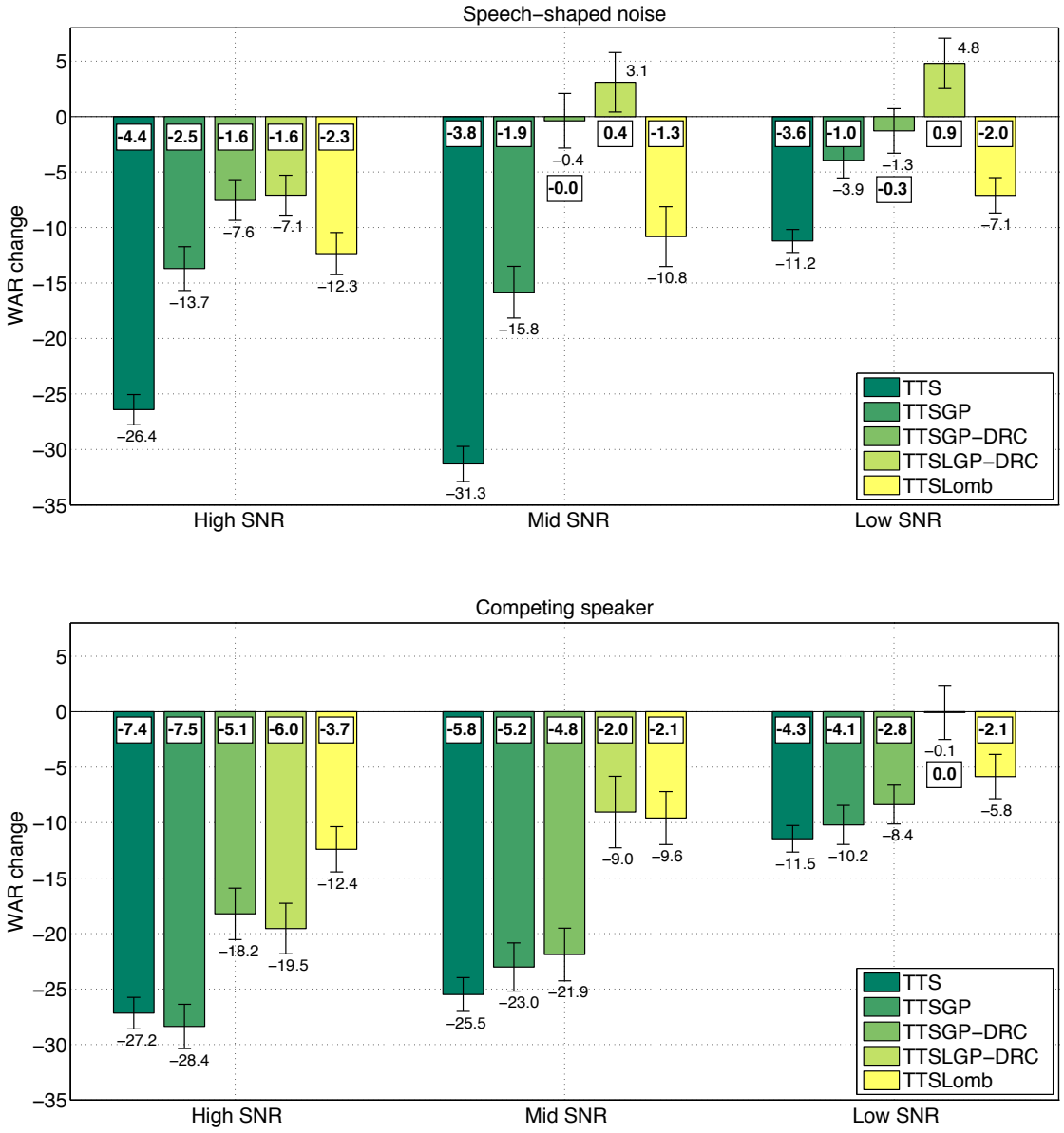


Figure 7.12: WAR change and EIC relative to natural speech (value inside boxes) for speech-shaped noise (top) and competing speaker (bottom). The results for TTSGP and TTSLomb were obtained from Evaluation I, TTSGP-DRC from Evaluation II and TTSLGP-DRC Evaluation III.

7.6 Conclusions

In this chapter, we presented the results of three large scale listening experiments comparing the GP-based Mel cepstral modification proposed in the previous chapter to other intelligibility enhancement methods. The three listening evaluations tested intelligibility in two different maskers: speech-shaped noise (SSN) and competing speaker (CS). Each masker was presented at three different SNR levels.

In the first evaluation, we compared the GP method to HMM-adaptation of Lombard speech (TTSLomb). In this experiment, we also evaluated two natural voices, one plain read-speech (Normal) and other read but noise-induced speech (Lombard). The results show the large gap between intelligibility in noise of natural speech and an unmodified TTS voice. The gap was made smaller by GP modification and adaptation, particularly in SSN. Adaptation was better in the CS scenario showing the importance of changes in duration and excitation signal.

In the second evaluation, we compared the GP method to two noise-independent approaches – a spectral shaper and a dynamic range compressor (TTS-SS-DRC and TTS-SSE-DRC) – and to one noise-dependent method also based on a speech intelligibility objective measure – the speech intelligibility measure (TTS-OptSII). These three methods were originally proposed for natural speech but in this evaluation were used to post process TTS generated speech. We also evaluated two method combinations (TTSGP-DRC and TTSGP-SS-DRC) and natural plain speech. Although the methods shared similar spectral gain shapes, the absolute gains and their modal natures – the shape of the boosting and the number of boosted regions – were quite different. The most effective strategy in SSN was TTS-SS-DRC, a noise-independent unimodal spectral gain combined with DRC. We also observed that some styles were more intelligible than natural speech, which shows how effective these methods can be when applied to a synthetic voice. In the CS masker, all methods performed significantly worse than natural unmodified speech, except for the lower SNR condition where TTS-SS-DRC obtained intelligibility scores similar to natural speech. Although noise-dependency did not give any additional gains, the difference in intelligibility results obtained for the two maskers motivates the use of noise-dependent methods.

Our third evaluation, compared the TTS unmodified baseline and natural speech plain baseline to a combination of enhancement strategies, each of which was shown to be individually successful: the perceptually-motivated spectral shaper based on the Glimpse Proportion measure proposed in the previous chapter, dynamic range com-

pression, and adaptation to Lombard excitation and duration patterns. With this combination we achieved substantial intelligibility improvements relative to unmodified synthetic speech: 4.9 dB in competing speaker and 4.1 dB in speech-shaped noise. An analysis conducted across this and the other two evaluations shows that the spectral shaper and the compressor (both of which are loudness boosters) contribute most under higher SNR conditions, particularly for speech-shaped noise. Duration and excitation Lombard-adapted changes are more beneficial in lower SNR conditions, and for the competing speaker noise.

By combining duration and excitation changes with other techniques, we have managed to increase intelligibility for competing speaker noise to a level comparable to that already obtained for stationary noise. Although the Lombard changes made by the natural talker were induced by a mismatched masker, adaptation to Lombard duration and excitation models contributed to gains not only in the competing speaker but also for the stationary masker at the lowest SNR. This approach however still entails the use of recorded Lombard speech of that particular speaker.

We mentioned in Section 2.1 that several mechanisms are involved in the perception of speech in noise: auditory grouping, glimpsing, linguistic-driven adjustments and spatial and visual cues. Auditory grouping and glimpsing take place in the low level peripheral auditory process and deal with the effect of energetic masking. The linguistic knowledge driven mechanism takes place in higher levels of processing, relying on a more abstract representation of speech. In the next chapter, we will present studies on how to use some higher-level information – the confusability of a word – to enhance intelligibility of synthetic speech in noise.

Chapter 8

Using top-down information

Thus far, to enhance intelligibility of TTS voices in noise, we have mainly focused on minimizing the energetic masking effect on the peripheral auditory system. We have shown that it is possible to get substantial intelligibility improvements by modifying spectral and duration components in such a way that more glimpses of speech appear in noise. The process of hearing in noise however involves other mechanisms. As discussed in Section 2.1, top-down information is also used by listeners to aid intelligibility in noise. As opposed to the information provided by the acoustic scene (bottom-up), the top-down information can provide cues that are derived from higher and more abstract representations of speech stored in the brain: language. This linguistic information can provide a prior to the listener that will help recognize information in adverse conditions. For instance, the expectation that spoken sentences are congruent, sensible, and grammatically correct, aids the intelligibility of speech as it limits the number of possible interpretations, making the decoding process easier. Linguistic information also manifests itself by the existence of a basic finite set of spoken units: the words. The lexicon of a listener can help the hearing task by providing a further constraint to the decoding process. There is a wealth of literature on the influence of words in the perception of speech in adverse listening conditions. Inspired by this literature, we exploit the use of top-down information provided by words to increase intelligibility of synthetic speech in noise. Part of the work presented here was published in Valentini-Botinhao et al. (2013b).

8.1 Spoken word recognition

Successful communication depends on word recognition, the process where a highly variable signal like speech is mapped to a discrete set of representations, the words. In a sentence or in isolation in either clean or steady noise conditions, words are not equally confusable. This inherent property of words interacts with factors like the communication channel (noise) and the context: different confusions arise in different noises and contexts. As all these factors interact with each other, predicting word confusability is not an easy task. Modelling word confusability is a much broader concept than modelling word intelligibility as it aims to provide not only a prediction for how intelligible a word is but also the possible confusions that can arise from hearing a word in a particular context and acoustic scene.

The objective measures for intelligibility evaluated in Chapter 4 provide predictions that are highly correlated with subjective scores of intelligibility at a word-level (is the transcription correct or not) when averaged across speakers, listeners, words and context. This shows how these measures are robust to different listening scenarios. However, as we will see in this chapter, when correlations are taken across all datapoints, including the effect of words and context, predictions are much poorer. In a nutshell, a complete model of word confusability in a sentence needs to consider what we refer to here as *acoustic confusability* (how similar words are compared with other words in the lexicon of the listener) and *linguistic confusability* (how likely a word is to be spoken in a particular context).

There are many studies investigating what influences spoken-word recognition. It is beyond the scope of this work to introduce them in great detail. A good summary of spoken-word recognition theories can be found in McQueen (2007) and a discussion of a model for spoken-word recognition and their limitations can be found in Luce and McLennan (2008) . McQueen (2007) mentions the three following units as sources of information provided to the process of word recognition: segmental (phone level), suprasegmental (syllable and word stress) and word-form (word frequency and phonology). To account for these sources on the process of word activation and competition Luce and Pisoni (1998) proposed the neighbourhood activation model. This model claims that in a listening task a group of words will be activated by a spoken word when they are similar to the target. The set of neighbouring words defines what they call a similarity neighbourhood, which is affected by two factors: the neighbourhood density (ND) – number of words in the neighbourhood – and the neighbourhood fre-

quency – how frequent these words are. They provide a formula to account for recognition rate as the ratio of a word's own frequency to the weighted sum of this value plus the frequency of its neighbours, accounting for the frequency of the word on its own as the basis of linguistic confusability. This means that words with more neighbours and with neighbours whose frequencies are higher are harder to recognize. Words with fewer neighbours and with neighbours that are not as frequent are easier to understand or recognize. Neighbours are weighted by phonetic distances obtained with confusion matrices derived from listening experiments. Other definitions of activation have been proposed based on the dynamics of lexical competition, which entails the impact over time of word frequency (facilitatory), onset density – words that begin with the same syllable or phone(s)– (inhibitory) and neighbourhood density (Magnuson et al., 2007).

Lexical complexity, that is how complex the word structure is, also influences the intelligibility of a spoken word. In adverse conditions such as additive noise we can expect that morphologically complex words (longer duration, more syllables) might be easier to recognize as their complexity provides more structural cues for the listeners. Francis and Nusbaum (1999) looked into the effects of lexical complexity on the intelligibility of natural and Text-To-Speech generated by the Votrax Type-n-Talk and DECTalk systems. They found that that morphological complexity does not aid intelligibility in quiet of the lower quality TTS stimuli, to the contrary of what was observed for natural speech. The authors suggest this could be due to the fact that low quality TTS systems provide poor segmental structure, so more complex word structures can become harder to recognize.

One can also look at word confusability as a source of listening adversity which in turn can influence how words are pronounced. To find whether we pronounce highly confusable words more clearly, similar to studies in Lombard speech, studies with highly confusable words have also observed that different acoustic patterns arise in words that are highly confusable, showing the intent of speakers to overcome this adversity. A study on the effect of lexical frequency and the Lombard reflex in Cantonese (Zhao and Jurafsky, 2009) found that both word frequency and noise influence tonal production: speech produced in noise showed higher F_0 contours as did the lower frequency words produced in quiet. Lower frequency words also presented wider F_0 range (tone dispersion) which was not seen for Lombard speech. If talkers, under the physical constrains of language production, make a specific effort towards wishing to be more understood, we would expect confusable words to be pronounced in a hyperarticulated fashion. In Buz and Jaeger (2012) we see a discussion on the matter of du-

ration and vowel space changes. For isolated word and scripted sentence production, phonological confusability, approximated using the phonological neighbourhood density model described previously, results in greater vowel dispersion and longer spoken word duration (Scarborough, 2010). In unscripted speech however the effect observed was in fact the opposite: highly confusable words were shorter and presented a smaller vowel space (Yao, 2011; Gahl et al., 2012). Buz and Jaeger (2012) claim this conflict comes from the fact that previous studies were not looking into contextual confusability, a factor that also affects speech recognition and speech production as their results seem to indicate. The influence of linguistic content on the Lombard effect has been investigated in Patel and Schell (2008). The authors wanted to investigate whether information bearing words (content words) are more enhanced when produced in noise compared to function words. They observed this effect only in the more adverse conditions, higher levels of noise. Words that referred to agents, objects and locations were prolonged and in some cases further enhanced by an increase in F_0. At moderate noise levels the modifications were more uniform across different word classes. This motivates speech modification strategies that take into account word-level confusability.

8.2 Using word confusability to increase intelligibility

Some words are inherently more intelligible than others i.e., they are less likely to be confused with other words. This property of words is currently ignored by intelligibility enhancing methods – none of the twenty entries of the Hurricane challenge used this (Cooke et al., 2013) – but it could potentially be exploited when applying speech modifications to TTS. The premise is that modifications aimed at improving intelligibility should not necessarily be "on" the whole time. For example, we can think in terms of energy budget whereby energy in a sentence is reallocated on the basis of the expected intelligibility of a word. More or less energy is expended depending on the predicted intelligibility of a word. Additionally, word confusability can be used to control the balance between naturalness/quality degradation and intelligibility improvements that modifications create. A modification can be seen as a deviation from natural sounding speech and might introduce unnecessary distortions. In this case, the level of modification could be constrained by the degree of distortion it introduces and the word confusability.

This chapter is a first attempt towards making use of a model of spoken word activation, the neighbourhood density model, in an energy-based speech modification.

For that we address two questions: how reliably can we predict intelligibility at a word-level and can we use this information to improve overall word recognition by selectively boosting highly confusable words.

There are many factors that influence the intelligibility of a word: acoustic confusability, linguistic confusability, the inherent intelligibility of a speaker, environmental factors (e.g., noise types) and listener characteristics. How to predict which words in a sentence are going to be easily intelligible and which ones are harder is not straightforward. Our previous intelligibility enhancement methods did not consider word-level information and in order to make the use of this information in a measurable task, we decided on the following constraints: we only consider word-level acoustic confusability (no linguistic confusability), we focus on synthetic speech from one speaker and on one type of noise (speech-shaped noise). As said previously word-level intelligibility can be used to control the amount of modification serving as a balance between the increase in intelligibility and the deviation from plain speech production. It is not clear however how to measure this deviation and more importantly what is an acceptable level of distortion. In order to constrain the amount of modification we adopt instead the SNR, a clearly defined value. As done in previous chapters we constrain the modification to a fixed sentence level SNR. The modification we are looking at then is energy reallocation across words, which we view as a starting point for other types of modifications.

The remainder of this chapter describes the work through the design and results of three listening tests. In the first experiment, we evaluate intelligibility of isolated words in noise. For this, we select words according to their neighbourhood density and frequency. In the following two experiments, these words are placed in matrix style sentences and we evaluate different energy reallocation strategies based on the intelligibility scores obtained in the isolated word experiment. In the second experiment, we evaluate intelligibility improvements of boosting one word using the energy from another word in the sentence and in the third experiment boosting one word with energy taken from all other words in the sentence. The following sections describe the set-up of the experiments, our findings and a discussion of the results.

8.3 Measuring word confusability

In Chapter 4 we evaluated a set of objective measures against subjective scores of intelligibility. Although we found measures that are highly correlated with subjective

scores we present here a possible set of objective measures that can be used for a more challenging task: measuring word-level confusability.

8.3.1 Neighbourhood density

Lexical or phonological neighbourhood density (ND) plays an important role in word recognition. Words with many lexical neighbours, differing by one phoneme insertion, deletion or substitution are more difficult to recognise than words with few lexical neighbours (Luce and Pisoni, 1998). In Cara and Goswami (2002) a second definition of phonological neighbourhood is given: the OVC-metric. In this metric, words that differ by insertions, deletions or substitutions in either the onset, vowel or coda of a word are counted. Words like *main* and *strain* are phonological neighbours in addition to for example *main* and *gain* according to the OVC-metric.

8.3.2 HMM-based distance

Measuring word confusability automatically using statistical models in the context of automatic speech recognition (ASR) is a way to analyse word error rate variability across datasets. In this context, studies focus on predicting from text how well an ASR system might perform. Although a different task from predicting human speech recognition, this can give us some insight as to how one can use a statistical model for word-level confusability measurement.

Bouwman et al. (2004) proposes to predict ASR word correct rate from what they call acoustic and linguistic confusability. A word A is considered to be acoustically confusable when its feature representation causes acoustic models of word B, where $B \neq A$, to produce likelihood scores that are equal to or higher than the score of the word's own model. Linguistic confusability follows a similar concept but concerns the likelihood scores assigned by the language model. To measure acoustic confusability (AC) between two words Bouwman et al. (2004) use a Kullback-Leibler (KL) divergence measure between the probability density functions of two different word models. The set of possible confusable words is limited by a fixed amount (ten) nearest neighbours and the overall AC score is given by the log of the average of the exponentiated divergence scores. The authors measure linguistic confusability (LC) using a bigram language model. They create a predictor for word correct rate (WCR) by clustering the two dimensional space covered by AC and LC measures calculated on a dataset into 36 regions and calculating the average WCR for each region. Testing on another dataset

they found that for on average 94% of cluster points WCR did not differ significantly from the predicted cluster value.

We can transpose this idea into the problem of measuring word confusability of speech generated by statistical models. In that sense we are measuring confusability arising from the speaker rather than confusability originating in the listener. We can predict for instance how well a TTS voice can distinguishably pronounce neighbouring words by looking into the distance between the models that generate these words. For that we look into the following distance measure between two observation vectors of different lengths (Juang and Rabiner, 1985):

$$D_S(\lambda,\lambda_i) = \frac{D(\lambda,\lambda_i) + D(\lambda_i,\lambda)}{2} \tag{8.1}$$

where

$$\begin{aligned} D(\lambda,\lambda_i) &= \frac{1}{T}\left[\log P(\boldsymbol{O}|\lambda) - \log P(\boldsymbol{O}|\lambda_i)\right] \\ D(\lambda_i,\lambda) &= \frac{1}{T_i}\left[\log P(\boldsymbol{O}^{(i)}|\lambda_i) - \log P(\boldsymbol{O}^{(i)}|\lambda)\right] \end{aligned} \tag{8.2}$$

where λ is the model that generated the target word, λ_i the model that generates the neighbour i, $\boldsymbol{O}$ and $\boldsymbol{O}^{(i)}$ are the observation vectors generated by models λ and λ_i respectively of length T and T_i.

To calculate confusability using this measure we can define the following value:

$$C(\lambda) = \sum_{i=1}^{N} \frac{1}{D_S(\lambda,\lambda_i)} \tag{8.3}$$

If the index i is covering the N size neighbourhood as defined by the OVC metric, then this measure is a weighted sum of OVC neighbours, where the weight is set to be the inverse of the distance between the neighbour and the target word. That is, the further the neighbour is from the target word the less it contributes to the confusability of the target word.

In the context of providing weights for the ND calculation we can see the HMM-based distance as a speaker dependent objective alternative for confusion matrices that are obtained through listening experiments as done in Luce and Pisoni (1998). An illustration of how this distance works is shown in Fig. 8.1, where we present a two dimensional Sammon mapping, a type of multidimensional scaling analysis, done on the distance matrix built for the verbs *sell*, *buy* and *see* and two coda neighbours (share the same ending). We obtained the distance matrix by calculating the HMM-based distance given by Eq.(8.1) between the models of each of these words (matrix of dimension 9x9). To calculate the distance between word described by the model sequence λ

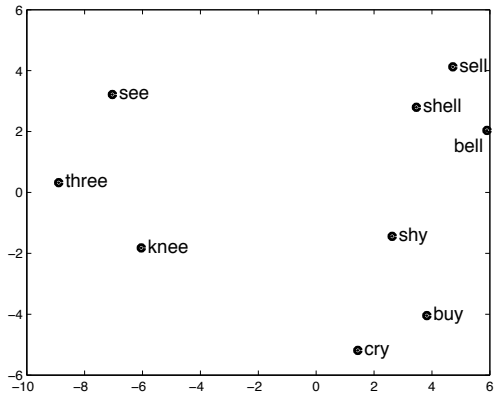


Figure 8.1: Multidimensional scaling of the HMM-distance matrix built for three verbs and their neighbours via a two dimensional Sammon mapping.

and λ_i we first generate observation vectors $\boldsymbol{O}$ and $\boldsymbol{O}^{(i)}$ from these models. We then calculate the likelihoods of Eq.(8.2). The synthesis models used for this calculation are the ones that will be used for the experiments of this chapter. A description of how it was trained and how we synthesized isolated words from it will be given in Sections 8.4 and 8.5.2. We can see in Fig. 8.1 that the dispersion within neighbours, like the words *three*, *see* and *knee*, is smaller than between different neighbourhoods. Additionally, words like *shell* and *shy*, that are onset neighbours, are closer to each other than words that share neither onset or coda, such as the words *shell* and *buy*.

8.4 Stimuli material

The synthetic voice we use in this experiment was built from plain read speech in the same manner as the TTS voice described in the previous chapter. For details on how this TTS voice was built see Section 6.4.1. As opposed to the experiments described in the previous chapter here we do not downsample the generated speech, instead we use the high quality generated signals sampled at 48 kHz.

To use as masker for the three experiments described in this chapter we generated speech-shaped noise from recordings of a female speaker sampled at 48 kHz, the same process that generated the speech-shaped noise used in our previous experiments (Cooke et al., 2012). As we decided to carry out the listening experiments with higher quality signals – sampled at a higher rate – this speech-shaped noise was obtained from the recordings of the female speaker sampled at 48 kHz rather than 16 kHz.

8.5 Isolated word experiment

The main goal of this work is to investigate whether we can improve enhancement methods by using additional information about word intelligibility. It is not clear however how to obtain this information and whether this can be obtained just from the words or the acoustics. In order to find this information we perform first a listening experiment with the words whose intelligibility we need to know. These subjective word intelligibility scores can be used to test whether signal based measures like the glimpse proportion, described in the Chapter 5 and the HMM-based measure described in this chapter are good predictors for word-level intelligibility. Also we can observe how much the neighbourhood density of a word affects its intelligibility in isolation.

8.5.1 Word selection criteria

As we are interested in investigating how to use a model of word-level intelligibility to improve the intelligibility of words in a sentence, to explain the selection of words we must first explain the design of the sentences. To control duration and complexity the desired sentences should have a fixed structure: same number of words and same syntactic organization. Additionally because we are investigating acoustic confusability, each word in the desired sentence should be as semantically predictable as the others. That is, the context is a factor but it affects all words equally. The appropriate set of sentences for this experiment is very similar to the sentences we used in Chapter 4, the Matrix sentences (Dreschler, 2006). Matrix sentences have a fixed syntactic structure: [name] [verb] [numeral] [adjective] [noun], for example: “Rachel has 5 blue rings”. To fill in each syntactic slot a word is chosen from a set of 10 words.

For our experiments we created Matrix-style sentences of a slightly different form: [imperative verb] the [adjective] [adjective] [noun], for example “Buy the French new car”. The choice of a different form was mainly due to the coverage of the monosyllabic lexical database (Cara and Goswami, 2002) used for the word selection which did not include proper names and conjugated verbs. The numeral slot was replaced by an adjective slot as there are not enough monosyllabic numbers to create a representative range of neighbourhood density values.

The word selection was done from an existing monosyllabic lexical database (Cara and Goswami, 2002). We chose 10 verbs, 20 adjectives – 10 for each sentence slot – and 10 nouns. This lexical database, provided in Cara and Goswami (2002), contains both neighbourhood density and frequency statistics which allowed us to select words

based on these values. Our first selection criteria, similar to the one used by Cooke (2009), was to only choose words whose frequency in written and spoken language is above 10 appearances per million. We are interested in looking into the effect of acoustic confusability only and to minimize the effect of a word spoken and written frequency we choose words that are familiar enough. As we were interested in covering a wide range of acoustic confusability our second criterion was to chose, for each syntactic category, words that fell into the two following categories:

- **hard** - words from a dense neighbourhood, ND-OVC $\geq N_H$
- **easy** - words from a sparse neighbourhood, ND-OVC $\leq N_E$

Another criterion used in word selection was sentence congruency, that is we choose words that could potentially create congruent sentences. For that we selected nouns to be only objects which can be acted upon and adjectives that can be associated to most of the nouns. Under all these three constraints the resulting easy and hard ND-OVC threshold values were: $N_H = 37$ and $N_E = 17$. Fig. 8.2 shows the histogram distribution of the ND-OVC for verbs (top), adjectives (middle) and nouns (bottom), where easy and hard words are represented by blue and red bars respectively. Note that the interval that separates easy and hard words is not the same across the different syntactic categories. For the list of words that we selected and their classification (easy or hard) refer to Table D.1 in Appendix D.

8.5.2 Word stimuli generation

As the TTS models were trained with sentences rather than isolated words, to obtain isolated words from the TTS voice we synthesize speech from carrier sentences of the format: “Now we will say “pause” *word* “pause” again”. We included the pauses to minimize coarticulation between the target word and the surrounding words which in turn minimizes segmentation artefacts. The target words were automatically segmented from the carrier sentence and added to noise with 200 ms initial and final lags.

8.5.3 Listening experiment design

To find which words can benefit from being presented at higher SNR levels we need to obtain word intelligibility scores at a range of different SNR values. The SNR levels were chosen in a separate listening experiment that involved 10 participants. We chose

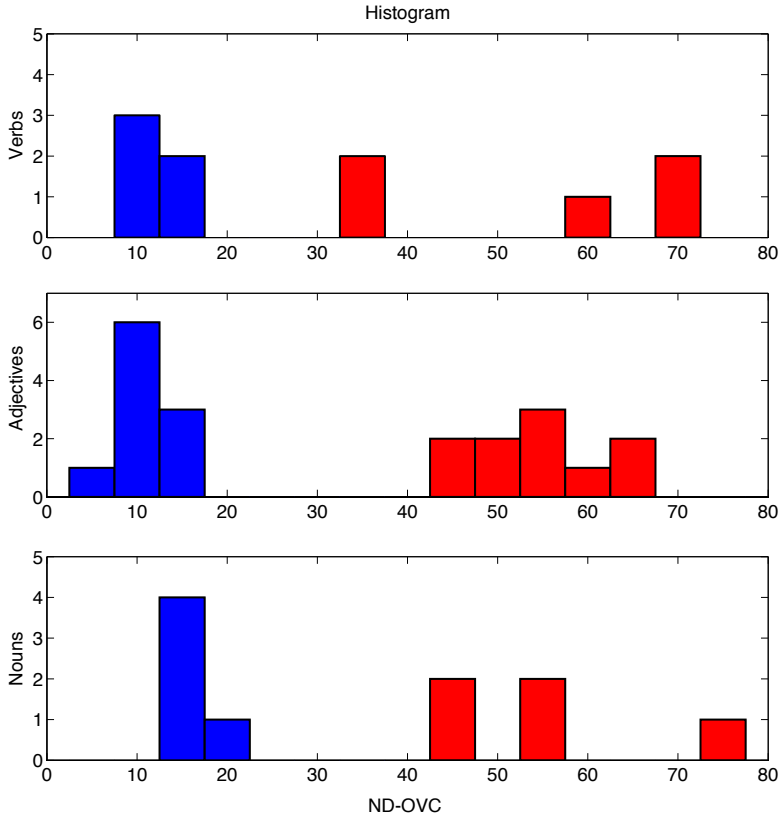


Figure 8.2: Histogram of the ND-OVC spread of the selected words: verbs (top), adjectives (middle) and nouns (bottom). Easy words are represented by the blue bars and hard words by the red bars.

a range so that in the highest SNR words classified as hard were intelligible and in the lowest SNR easy words were not intelligible, which resulted in the SNR range of -8 to 10 dB. Five SNR values were chosen from that range: -8, -3.5, 1, 5.5 and 10 dB.

For the actual test we had 25 native British English speakers with no hearing problem transcribing the words played in noise over headphones in sound-isolated rooms. Before the test, participants had to go through a practice session which included 20 other words synthesized in a similar manner as described in the last section and presented at the middle range SNR. After hearing all words presented at the five different SNRs, starting from the lowest to the highest SNR divided into four blocks, participants were also asked to transcribe the words in clean conditions, i.e. no speech-shaped noise present. Every participant heard the same stimuli (words and noise condition) but in a different order.

8.5.4 Results

In this section we present word accuracy rate (WAR) results calculated as the percentage correct word transcriptions across the 25 listeners. Figs. 8.3(a) and 8.3(b) show scatter plots of ND (ND-OVC) values versus the WAR obtained for each word in clean and in one of the noisy conditions (SNR= 5.5 dB). These results show that even in clean conditions a number of words were poorly understood, obtaining less than 60% WAR. We can also see that these words are mostly from a dense neighbourhood, belonging to the "hard" category. Although the linear correlation between ND and WAR is quite low (-0.46 for the clean condition and -0.31 for the noisy condition) when comparing the scatter plots of clean and noisy conditions we can see that the "easy" words are more robust to noise. That is, the dispersion towards the low WAR region triggered by the presence of noise is smaller.

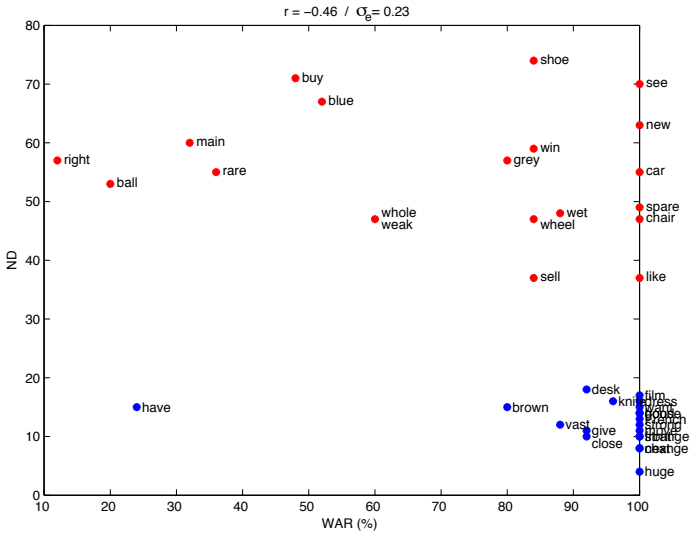
As each word was presented in noise at five different SNRs we are able to draw psychometric curves for each word. We present these curves in Fig. 8.4 for the verbs (top), nouns (middle) and adjectives (bottom). According to the NAM model we expect words classified as easy to have higher WAR than hard words. We can see however some words do not behave as expected. For instance the verb *have* classified as an easy word is in fact poorly recognized and that the verb *see* is easier to recognize than expected from its ND category. The easy and hard curves are less clearly distinguishable for the selected nouns. Easy words like *knife* and *dress* present quite a steep slope and are quite unintelligible in low SNRs, while the easy word *film* presents quite a flat and low WAR curve. The word *chair*, although a hard word, was the most intelligible noun across the SNR range. The hard and easy categories are more clear when looking at the psychometric curves of the adjectives, with fewer cases of words where the ND category expectations and the psychometric curve values are mismatched. A clear mismatch is the word *vast* that presents a WAR curve much lower than expected.

Although it is out of the scope of this work to explain why the ND categories of easy and hard did not perfectly match the intelligibility scores obtained in this isolated word experiment we comment on what we think may have contributed to this mismatch. Linguistic context and coarticulation, two factors that were not considered in the classification of easy and hard words, did not contribute to the mismatch as this was a isolated word experiment. We expect however that not only the ND of words but also both speaker and the noise contribute to the WAR observed. As the speech used in our experiment is not natural but from a TTS voice it can be expected that

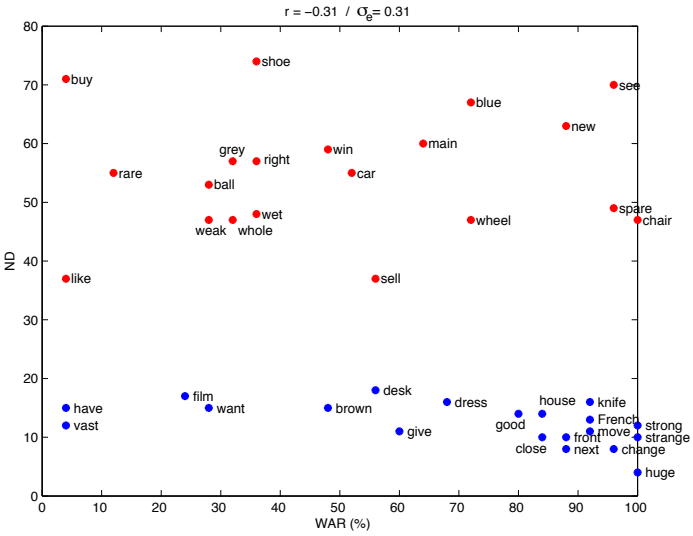
there is an effect of the quality of the synthetic speech. This can be seen in the WAR results in the clean condition, shown in Fig. 8.3(a). We saw that some words are very poorly transcribed even in the clean condition at a much lower than expected rate than seen for natural speech. For instance, in the experiment described in Luce and Pisoni (1998) only 36 out of 918 words (around 4 % of all words) were recognized at less than 90 % correct. In our experiment, 9 out of 40 words obtained less than 60 % WAR, representing 22.5 % of all words. Although the words in Luce and Pisoni's (1998) experiment and in ours were not the same, we think that the difference in recognition rates is notable. The speaker effect also presents itself in another way: words with few neighbours are more likely to be poorly generated by the TTS acoustic models. This is due to the fact that model accuracy is higher in fragments where more training data covering similar acoustic contexts is available in the training set. The opposite effect can arise then: easy words become hard words.

The noise on the other hand will interact with different words in distinct ways, affecting the possible word activations that form its neighbourhood. As the noise which the words were presented in is highly energetic in the formant region, we see that words that contain higher frequency components like *chair*, *see* and *spare* are more intelligible in noise than words composed mostly of vowel and nasal sounds. Word duration, which can be interpreted as an indication of lexical complexity, is another factor to consider as quite short words like *have*, *vast* and *film* although classified as easy words present a low intelligibility rate in our experiments.

To illustrate how challenging it is to "represent" intelligibility at the word-level we present in Fig. 8.5 predictions of word-level intelligibility using the glimpse proportion measure (GP) versus the subjective WAR results of all datapoints (40 words in 5 SNR). We showed in Chapter 4 that the GP measure obtained a high correlation coefficient (up to 0.94) with subjective intelligibility scores of a male TTS voice in diverse noise conditions when both GP and WAR scores were calculated at a word-level but averaged across the different words. Fig. 8.5 however shows a very poor correlation (0.44 correlation coefficient) between GP and WAR calculated for individual words. In turn, when both GP and WAR scores are averaged across words for each of the five different SNR conditions, values illustrated by the red dots in the figure, a much stronger correlation appears. Normalizing the GP measure with the ND value of each word improves the word-level correlation to 0.58, as presented in Fig. 8.6. By using the HMM-distance measure defined in Eq.(8.3) to calculate a weighted sum of the number of neighbours we can further improve this correlation to 0.62 as presented in Fig. 8.7.



(a) clean



(b) noisy: SNR=5.5 dB

Figure 8.3: Neighbourhood density versus word accuracy rates for a clean (a) and in noise (b). Easy words are represented in blue and hard words in red.

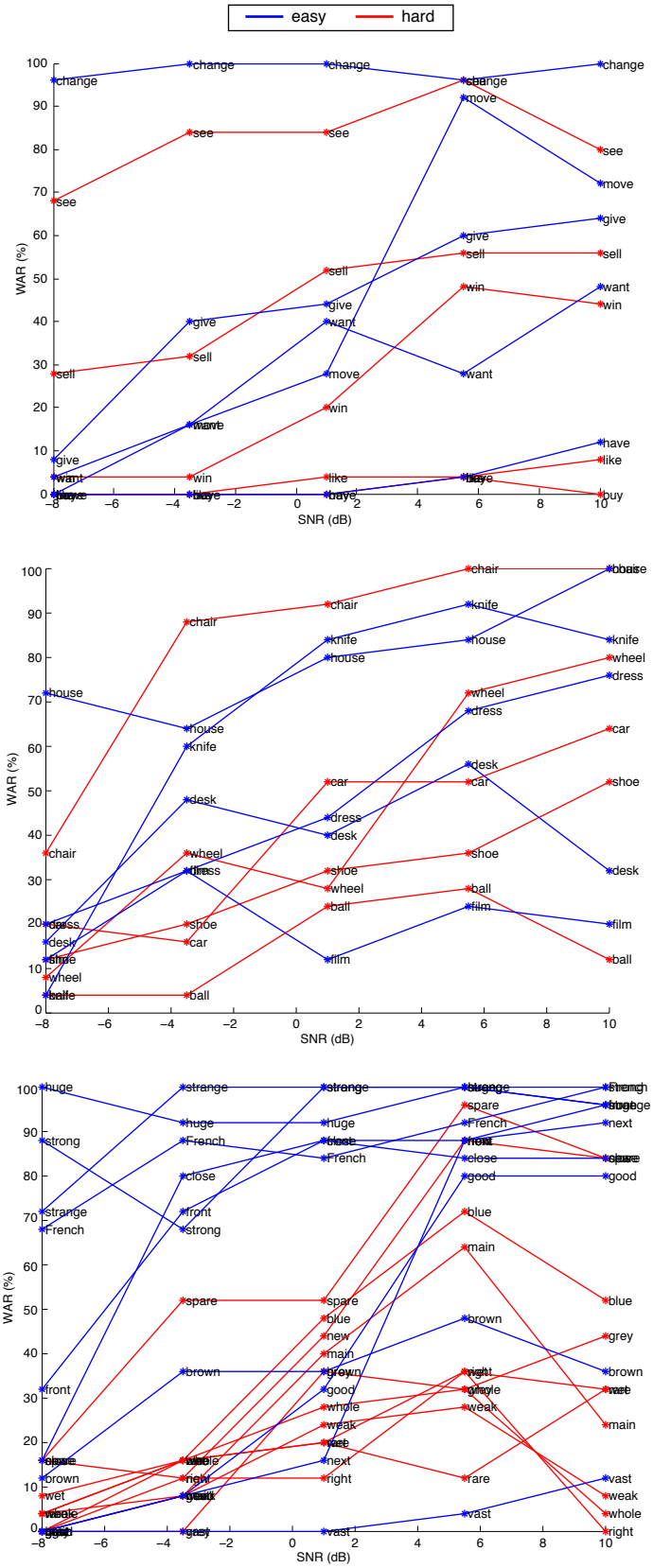


Figure 8.4: Psychometric curves of verbs (top), nouns (middle) and adjectives (bottom).

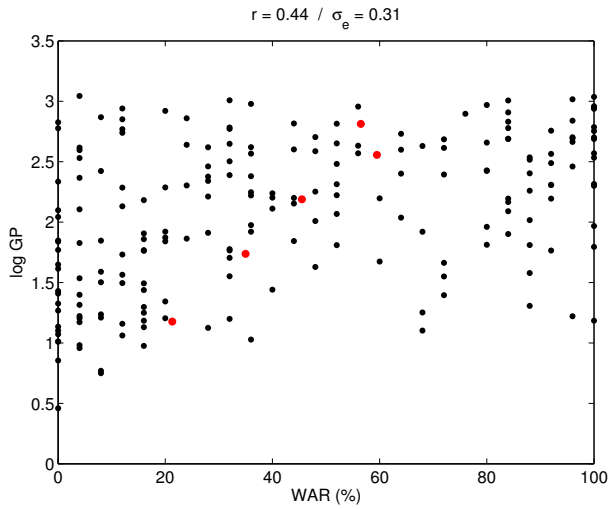


Figure 8.5: Scatter plot of GP and WAR values calculated at the word-level in all SNRs. Red points represent values averaged across words in each SNRs.

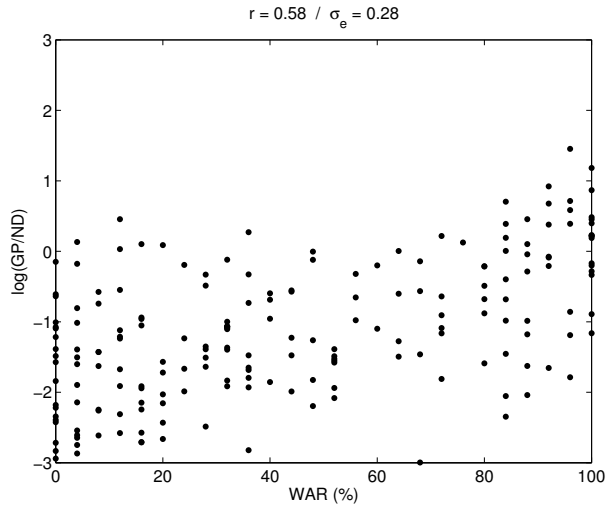


Figure 8.6: Scatter plot of GP values normalized by ND values and WAR values calculated at the word-level in all SNRs.

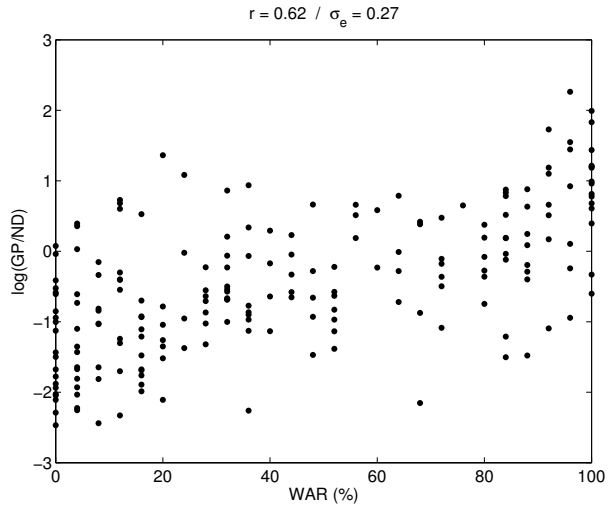


Figure 8.7: Scatter plot of GP values normalized by ND values weighted with a HMM-distance measure and WAR values calculated at the word-level in all SNRs.

8.6 Sentence experiment: giver and receiver

The main goal of this work is to find whether we can improve overall sentence intelligibility by using a prior on word intelligibility. To show that, we decided to selectively boost words in a sentence as a starting point for more complex modifications. We found from the previous evaluation which words from a pool of 40 words can benefit from increasing SNR, i.e. energy boosting. We did that by looking into their psychometric curves in isolation. Now we will use this information to derive energy boosting strategies for the sentence experiments. We start with the strategy of boosting a word – the receiver – with power taken from another word – the giver.

8.6.1 Giver and receiver proposed classification

For the sentence experiment we split the 40 words used in the previous experiment into two categories: giver and receiver. Because the ND, the GP and the HMM-based measures did not show a high correlation with the subjective data we decided to use the actual subjective scores as the classification criterion. A word was considered to be a giver if the WAR in isolation for all SNRs tested as either quite low – hard giver – or quite high – easy giver. Easy and hard now relate to their intelligibility scores rather than their ND values. That allows for two different strategies, attenuate words that are highly intelligible as they will not suffer as much from the attenuation. The second is to attenuate words that are too difficult to understand so they are not worth spending energy on. The receivers were words that had presented steep slopes in the isolation experiment, providing us evidence that at higher SNR values they were more intelligible. Our expectation was that they would benefit from energy boosting. Table D.2, in Appendix D, shows the list of words and their proposed giver/receiver classification. We present the psychometric curves averaged across receivers and givers (easy and hard) in Fig. 8.8. We can see that both giver curves are quite flat across the SNR range and that on average receivers are more sensitive to changes in SNR, which makes them promising candidates for selective boosting.

8.6.2 Sentence material

To create the sentence material we built Matrix-style sentences of the form: [imperative verb] [adjective] [adjective] [noun] so that in each sentence there is a word pair of giver/receiver. The other words in the sentence, referred to as fillers, were randomly

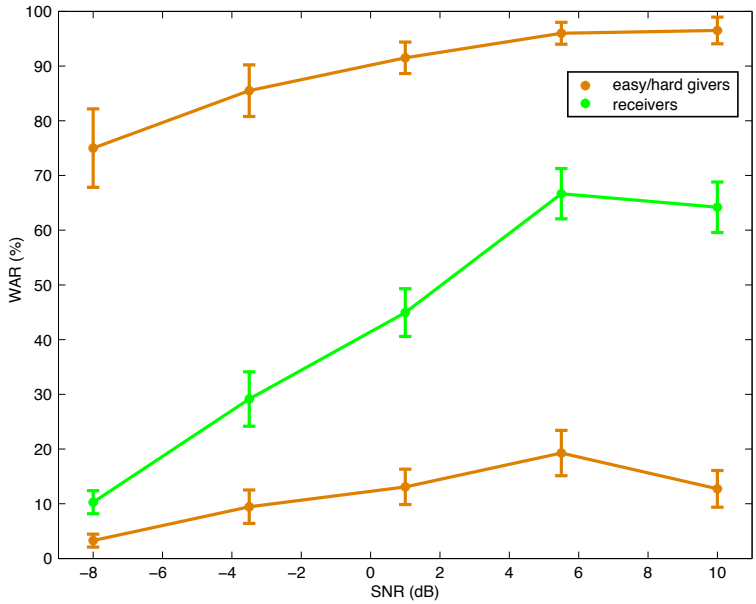


Figure 8.8: Psychometric curves of givers and receivers in isolation.

selected from the 40 word pool, so that each word does not appear more than six times in the whole set of sentences. As we had givers and receivers for all word categories, we could rule out, or analyse, the effect of the position of the giver/receiver words in the sentence by creating 12 different sentence groups. A sentence group is defined by the position of both the giver and the receiver in the sentence. The list of sentences and their sentence group is shown in Appendix D Table D.3 for the proposed word pair selection and in Table D.4 for a random word pair selection. A sentence group named G1_R2 refers to sentences where the giver is the first word in the sentence (verb) and the receiver in the second (adjective). Five different sentences were created for each group, resulting in 60 sentences in total.

Although we chose the nouns to be objects which can be acted upon, and adjectives attributes that can be associated to most of these nouns, we did not focus on designing sentences that made sense. This will also contribute to make the semantic context a smaller factor. The structure of the sentence, that is the fact that every sentence is of the structure [verb] [adjective] [adjective] [noun], will however be a considerable factor as it restricts the number of possible words that can fit in each slot. Participants were not made aware of this structure beforehand.

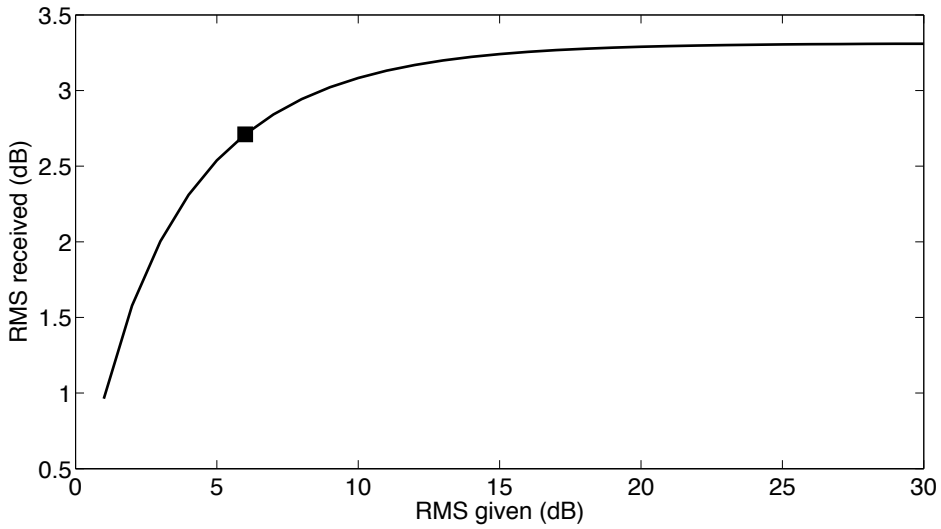


Figure 8.9: Curve of giver/receiver RMS exchange calculated for each word pair and averaged across all 60 pairs. The curve shows the relation where the RMS of the sentence is kept fixed. The black point illustrates the operation point chosen for our experiment: give 6 dB and receive 2.7 dB.

8.6.3 Modifications

To investigate whether selectively boosting the receiver word by attenuating the giver word can increase overall intelligibility we compared two types of modifications:

- **proposed** - select givers and receivers according to the results of the isolated word experiment, see classification criterion in Section 8.6.1;
- **random** - select pairs of givers/receivers randomly.

From the results of the previous experiment we expect that different receivers need different amounts of boosting to raise their WAR to a similar value, but to simplify the experiments we decided to fix the amount of power level loss to 6 dB. On average the receivers power increases to 2.7 dB with the constraint that the overall energy level of the sentence remains unchanged. The amount of power received by a word depends on the power values of both giver and receiver and on their durations, that is a shorter and quieter receiver will effectively be boosted more than a longer and highly energetic word. The power value being transferred depends on the giver: a long highly energetic giver will give more energy than a short and quieter one. To illustrate this, Fig. 8.9 shows how much in dB on average a word can receive from a certain amount of given power value so that the overall power of both words is fixed. The operational point

that we chose, illustrated by the black square, is a compromise between not losing too much energy and receiving enough energy: give 6 dB and receive 2.7 dB.

To change the power of the signal for the segments defining the giver and receiver words first we calculate where the word starts and finish by using the state level duration information generated by the TTS models. Second, to change the power of the signal defined by these time stamps we apply a scale factor to the speech waveform. As we would like to keep the power of the sentence – or the energy since sentence duration is not modified – the following holds:

$$SNR = 10\log\left(\frac{P_S}{P_N}\right) \tag{8.4}$$

$$P_S = \sum_{t=1}^{T} s^2(t) \tag{8.5}$$

$$P_S = P_R + P_G + P_F \tag{8.6}$$

$$P_S = P'_R + P'_G + P_F \tag{8.7}$$

$$P_R + P_G = P'_R + P'_G \tag{8.8}$$

where P_S is the power of the whole sentence, P_N the power of the noise signal across the whole sentence, and P_R, P_G and P_F the power contained in time interval defining the receiver, the giver and the filler words. P'_G and P'_R are the modified power values of the giver and receiver word segments.

If the giver loses 6 dB its new power level P'_G is given by:

$$10\log P'_G = 10\log P_G - 6 \tag{8.9}$$

$$P'_G = 10^{(10\log P_G - 6)/10} \tag{8.10}$$

The scale factor β_G that needs to be applied to the giver word to result in the P'_G level is calculated in the following way:

$$P'_G = \sum_{t=T_{G,i}}^{T_{G,f}} (\beta_G s(t))^2 \tag{8.11}$$

$$= \beta_G^2 \sum_{t=T_{G,i}}^{T_{G,f}} s^2(t) \tag{8.12}$$

$$= \beta_G^2 P_G \tag{8.13}$$

$$\beta_G = \sqrt{\frac{P'_G}{P_G}} \tag{8.14}$$

$$= \sqrt{\frac{10^{(10\log P_G - 6)/10}}{P_G}} \tag{8.15}$$

where $T_{G,i}$ and $T_{G,f}$ define the initial and final time index that define the segment containing the giver.

From the giver new power P'_G we can define the receiver new power level P'_R as:

$$P'_R = P_R + P_G - P'_G \tag{8.16}$$

The scale factor β_R for the receiver is then given by:

$$P'_R = \beta_R^2 P_R \tag{8.17}$$

$$\beta_R = \sqrt{\frac{P'_R}{P_R}} \tag{8.18}$$

$$= \sqrt{\frac{P_R + P_G - P'_G}{P_R}} \tag{8.19}$$

To attenuate artefacts that can arise from scaling the waveform using a rectangular window, we applied a trapezoid window instead. The trapezoid starts with the beginning of the word – as given by the state duration sequence – and ends with the completion of the word – also given by the state duration. The initial transition segment – from scale factor 1 to the midsegment of the trapezoid – lasts for 5 ms, as does the final transition segment. The length of any given word in the sentence set we use for this experiment is between 165 ms and 475 ms.

8.6.4 Listening experiment design

As said previously we created a set of 60 sentences organized in 12 groups defined by the position of the giver/receiver pair in the sentence. The 60 sentences were evaluated at three different modifications. In order to ensure that no listener heard any sentence more than once we divided the experiment across three groups of listeners. Each listener heard all 60 sentences once and the modification type applied to each sentence was spread across the listeners so the whole test (180 stimuli) was covered by each group of 3 listeners.

To simplify the experiment, we tested the three conditions (two modifications and unmodified) at only one SNR level. Before the main test we carried out a small listening experiment with five participants to find the sentence SNR level that would lead to an average of 50 % WAR across the 60 unmodified sentences. The SNR found was -3 dB.

Prior to the actual experiments, participants had to take a practice session which involved 20 sentences of the same structure filled with other words. At the end of the

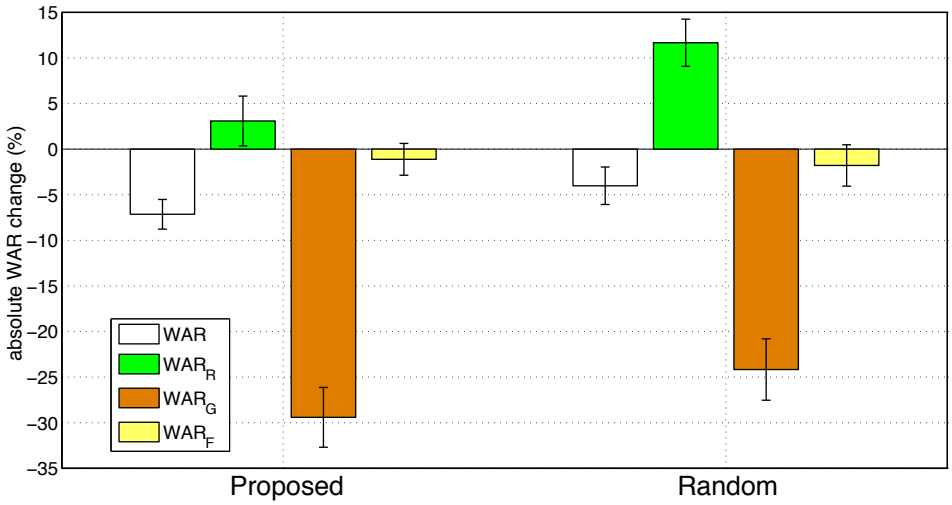


Figure 8.10: Absolute changes in WAR (in %) of proposed and random modification with respect to unmodified. Values for unmodified speech are: WAR = 49.6% and for proposed/random: $WAR_R = 51.25/53.3\%$, $WAR_G = 50.5/49.9\%$ and $WAR_F = 48.3/47.6\%$.

test all participants were also asked to transcribe the 60 sentences in clean. In total, 60 participants performed the test.

8.6.5 Results

We present the results averaged across words and listeners (deviation represents word dispersion only) for each modification in Fig. 8.10 in terms of absolute change compared to the unmodified case. The acronyms WAR, WAR_R, WAR_G and WAR_F refer to the intelligibility of all words, receivers, givers and fillers. As a reference, the rates obtained for unmodified speech were: WAR = 49.6% and for proposed/random: $WAR_R = 51.25/53.3\%$, $WAR_G = 50.5/49.9\%$ and $WAR_F = 48.3/47.6\%$.

We can see that for both modifications boosting a word at the detriment of another word decreases WAR results. The intelligibility of the givers drops significantly in both cases, more than a 25% absolute drop, while the receivers only gained up to 12% in word accuracy. The results also show that on average choosing the pairs randomly rather than according to the isolated word experiment generates more gains for receivers and smaller WAR drops for givers. The following sections present a more localized analysis of the differences between the scores for the isolated word and sentence experiments, the effect of the position in the sentence of the giver and receiver pairs and revisit the overall findings.

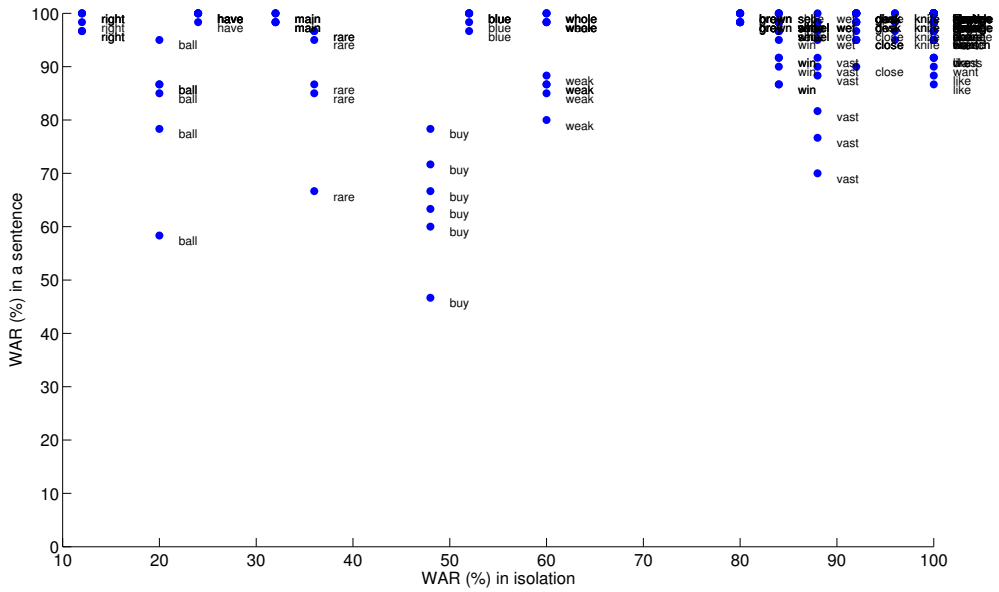


Figure 8.11: Word accuracy scores obtained in isolation and in a sentence in clean conditions. The dots refer to scores obtained for a word in different sentences.

	easy words	hard words
isolation	93.2 (3.8)	71.2 (6.5)
sentence	97.8 (0.8)	94.4 (2.0)

Table 8.1: WAR (%) and standard error of easy and hard words in isolation and in a sentence in clean conditions.

8.6.5.1 Words in isolation versus words in sentence

Comparing the WAR for each word obtained in the isolated word experiment and in the sentence experiment we can see the effect that co-articulation and context or sentence structure have on the intelligibility of receivers and givers. In this section, we present this comparison for the clean condition in Fig. 8.11 and the noisy condition in Figs. 8.12 and 8.13(a) at the word class level (easy givers, hard givers and receivers) and in Fig. 8.13(b) at the word-level.

Fig. 8.11 shows the averaged WAR obtained in the clean condition, in isolation and in the sentence experiment. No energy reallocation is applied to the sentences used in clean. We see up to six dots for the sentence results referring to scores obtained by the same word in different context, i.e. different sentences. We can see that in clean conditions and in a sentence most words can reach more than 80% accuracy and that words that we saw as being very hard to recognize like the words *right*, *have* and

main are highly intelligible in a sentence. On average across all six possible contexts, all words are more intelligible in a sentence than in isolation when they are presented without any noise. We present the WAR of easy and hard words in Table 8.1 obtained in isolation and in a sentence. Easy words are more intelligible in both scenarios but the difference is less pronounced in a sentence than in isolation. These results indicate that the effect of neighbourhood density on the intelligibility of words is limited when a word is presented with context.

To compare results across the two experiments in noise we first calculate the SNR at which each word was presented in the sentence experiment. This allows us to compare results for the same SNR, so that any difference in WAR is only due to the isolation/sentence effect. Although we set the sentence SNR to -3 dB it does not mean that each word in the sentence is presented at this SNR. We calculate for each sentence the SNR that the giver and receiver is presented at. As the same word appears as a giver (or receiver) in more than one sentence, we obtain the word SNR by averaging across its occurrences (as either a giver or receiver, filler is not being counted here). In the end we have one unique SNR per word. The same procedure is done to calculate the WAR of each of the 40 words in their functions as givers and receivers. The results are then averaged within each category: easy givers, hard givers and receivers. These values and their standard error (which represents the variance across the words in the category) are shown in Fig. 8.12. Note that the sentence results only contain two points along the x-axis (SNR), because words were either boosted (receivers) or attenuated (givers), whereas in the isolated word experiment words were played at five different SNR values.

We can see that, in general, words played in noise are also more intelligible in a sentence than in isolation: hard givers and receivers are more intelligible in a sentence than in isolation (WARs: 11% to 31% hard giver; 31% to 58% receiver). Easy givers are on average less intelligible in a sentence (WARs: 88% to 74%) and most of this drop comes from the words *huge*, *French* and *strange*.

The slopes of the curves are also different, that is, the easy and hard givers' WAR drops more than expected and receivers do not increase as much. It seems that too much energy is being taken from the giver and the receiver is still not getting enough, which is why the WAR rate per sentence does not increase.

To see how each word contributes to this difference we present in Fig.8.13(a) a scatter plot of WAR obtained in a sentence against the WAR in isolation, calculated by mapping the SNR presented at in the sentence to find the WAR using the psycho-

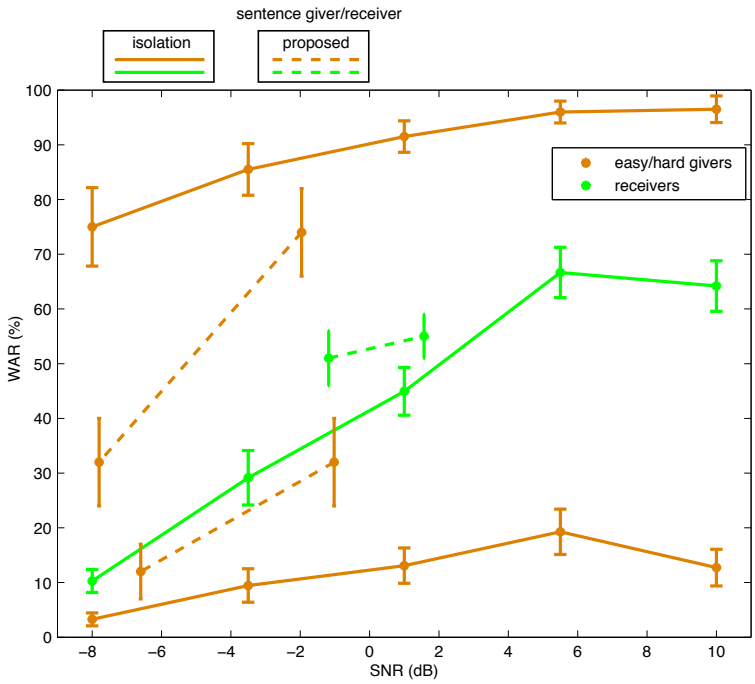
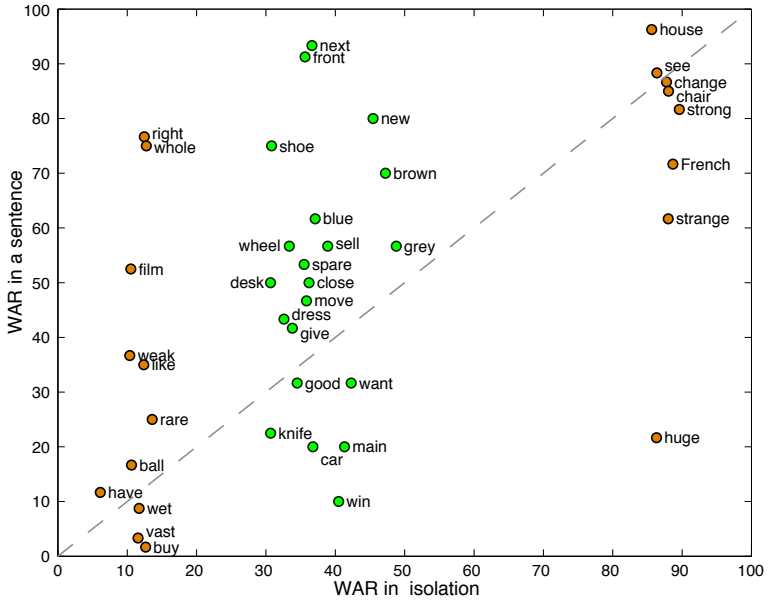


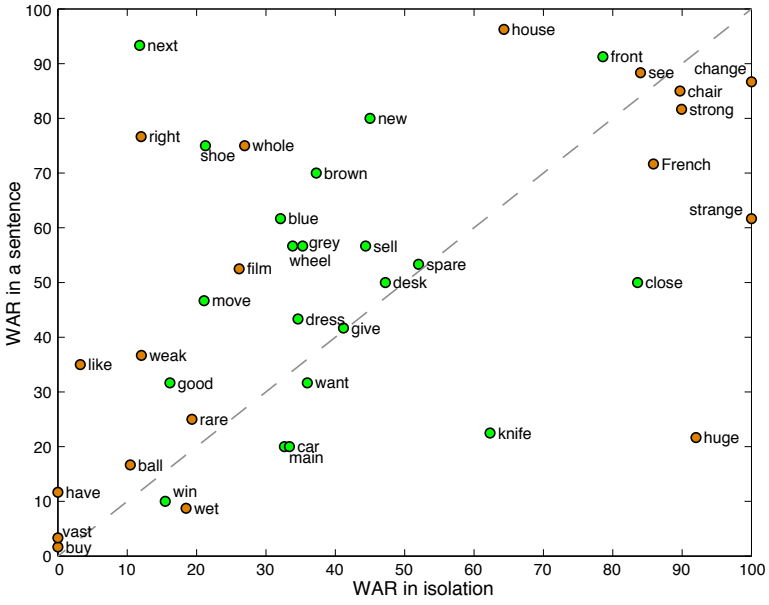
Figure 8.12: Psychometric curves of givers and receivers in isolation and in a sentence with the proposed modification (dashed).

metric curves of each word group – easy and hard givers and receiver. We can see that the group division does not transpose to the sentence experiment, as we see a lot of dispersion of WAR in a sentence within each group. When we map the sentence SNR to word individual psychometric curves obtained in isolation, see Fig. 8.13(b)) we can see that the effect of being in a sentence is more uniform across words but still not the same for all words: while most words are more intelligible in sentences some, particularly the easier words in isolation, became harder to understand in the sentence experiment.

Additionally we present in Fig. 8.14 the psychometric curves of randomly chosen givers and receivers. We can see that choosing giver and receiver pairs randomly brings their psychometric curves closer to each other. The givers lose less, as they are not mainly made of easy givers that we saw previously lose quite a lot. Although the SNR boost is quite similar to the proposed modification, receivers WAR increase significantly more.



(a)



(b)

Figure 8.13: Scatter plot of word-level accuracy rates obtained in isolation (by mapping each word to their group (a) or isolated word (b) psychometric curve obtained in exp1) and in a sentence (from exp2).

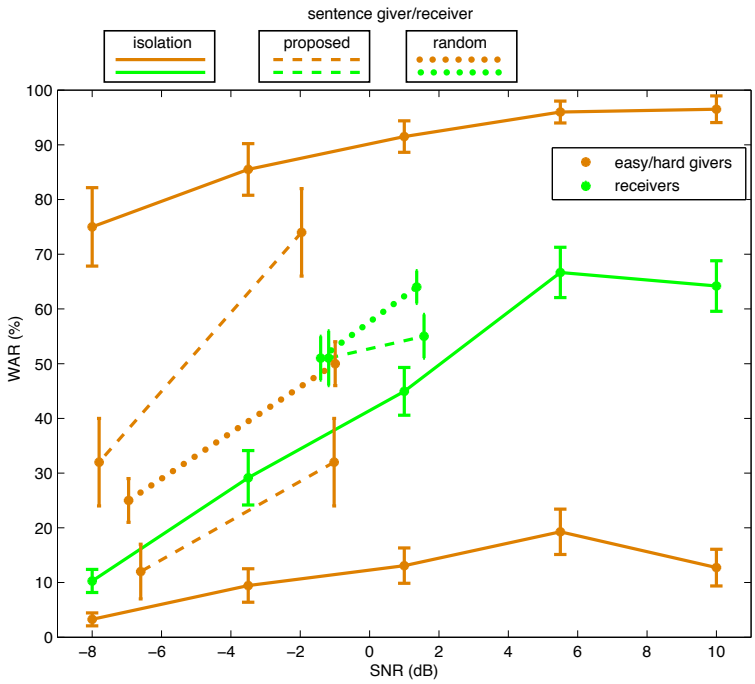


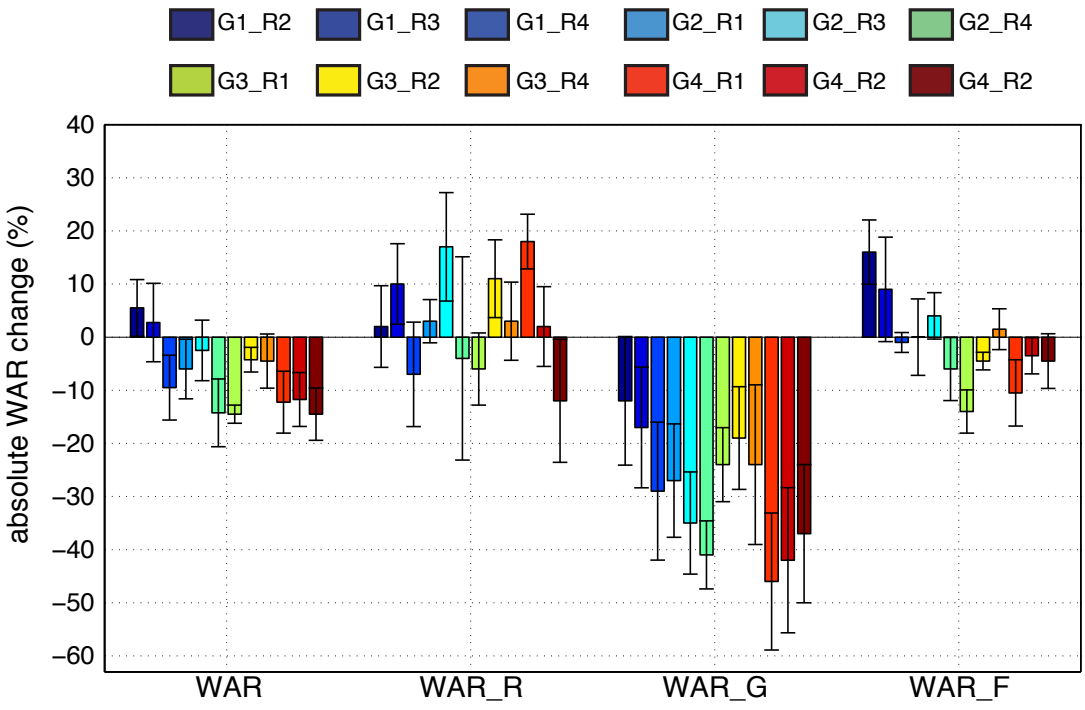
Figure 8.14: Psychometric curves of givers and receivers in isolation and in a sentence with the proposed modification (dashed) and the random modification (dotted).

8.6.5.2 Word pair position

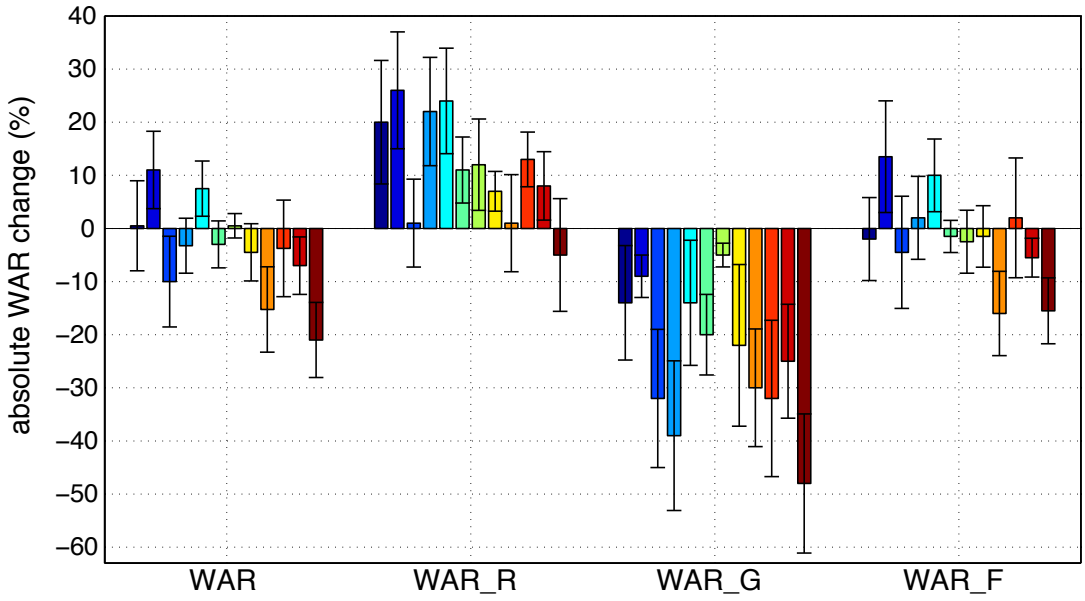
Although we saw in Fig. 8.10 that the intelligibility of filler words WAR_F, averaged across sentences, changes relatively little when we boost and attenuate pairs of words (from 48% to 47%), there is a large variation across sentences. To look into that we plot in Fig. 8.15(a) the relative WAR changes with respect to unmodified for the proposed modification separated into the 12 sentence groups as seen in Table D.3. Each group represents a different giver/receiver position in the sentence, their acronyms reflecting that – G1: giver position 1 and R3: receiver position three. Results for the random modification are presented in Fig. 8.15(b).

When we group the sentences into pair positions we see that the intelligibility of fillers changes more than 15% absolute WAR. As we do not have enough conditions in which the words chosen as fillers do not change we can not say how the position of the pair affects the fillers but we could see that even when we make localized changes other words in the sentence are also affected.

It is hard to tell what is the best position strategy but we can point out two strategies that did work for receivers: receiver that increase more are the verbs (groups: G1_R2, G2_R1 and G4_R1) and adjectives (groups: G1_R3, G2_R3 and G3_R2). Boosting the last word, the noun, did not bring any substantial improvements (groups with R4).



(a) proposed



(b) random

Figure 8.15: Relative changes in WAR (in %) for the proposed (top) and random (bottom) modification with respect to unmodified.

8.6.5.3 Overall findings

Isolation versus sentence:

- In clean conditions all words are more intelligible in a sentence than in isolation;
- giver/receiver psychometric curves in speech-shaped noise change significantly from isolation to sentence, both in terms of offset and slope;
- easy givers are less intelligible in a sentence than in isolation and see a larger WAR decrease than expected;
- hard givers are less intelligible in isolation and also lose intelligibility more than expected;
- boosted receivers on the other hand are not as intelligible as expected, in some cases presenting even a drop in intelligibility as we saw when boosting the nouns.

Modification strategy:

- Boosting RMS is not enough to increase receiver WAR above the losses of givers (for both proposed and random modification)
- Although we do not modify the SNR of the filler words we observed that their intelligibility changes;
- In a design where only one word gives to another, choosing the word that gives and the word that receives randomly generates a larger WAR gain – receivers – and less WAR drop – givers – than choosing according to their intelligibility in isolation.

8.7 Sentence experiment: boosting one word

We saw in the last experiment that boosting one word and attenuating another word from the same sentence impacts on the intelligibility of the other words in the sentence. Not only that, removing energy from one word to give to another does not improve overall intelligibility rates, mostly because the intelligibility of the attenuated word drops at a much higher rate than the boosted word increases. To overcome these two issues, we now try a different type of energy reallocation: we reallocate energy from the whole sentence to boost one word. In other words we emphasize one word in a

sentence while making the rest of the words quieter, which is also a naturally occurring modification. Although we saw in the previous experiment that randomly selecting givers and receivers resulted in higher receiver gains, for this experiment we still select receivers according to their scores in isolation (proposed selection). We decided to do so to be able to analyse the performance of the proposed selection under a more promising modification strategy and to see under which circumstances can this strategy improve scores.

8.7.1 Sentence material

In this experiment, the same sentence material is used as in the previous experiment, that is the 60 sentences displayed in Table D.3, Appendix D.

8.7.2 Modification

To check whether boosting one word in the sentence while keeping the overall SNR fixed can increase intelligibility we evaluate the following modifications:

- **Medium boost** - boost receiver and attenuate all other words in the sentence. Relative power level boost of 3 dB
- **High boost** - boost receiver and attenuate all other words in the sentence. Relative power level boost of 5 dB

In this experiment, we focus on giving the same boost across sentences and at two different levels of boosting. The boost values were chosen by fixing the amount of relative boost between receiver and the rest of the sentence to 3 dB and 5 dB, resulting in the Medium and High boost modifications. We can see these two operation points in Fig. 8.16 which represents the power levels received and given calculated for each sentence and averaged across the sentence set. The red line represents the relative gain between receiver and giver and the black/red dots the operation points chosen for this experiment. When comparing this figure to Fig. 8.9 we can see that for much lower values of attenuation it is possible to obtain similar RMS boosts when attenuation is carried out across the whole sentence rather than one word.

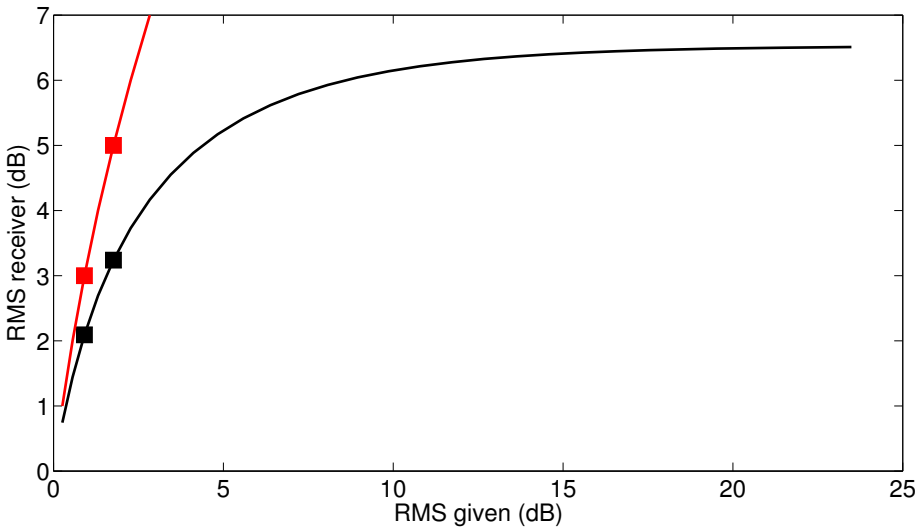


Figure 8.16: Curve of giver/receiver RMS exchange calculated for each pair and averaged across different pairs. The curve shows the relation where the overall RMS of the sentence is kept fixed. The points that we chose for our experiment: Medium boost give 0.9 dB and receive 2.1 dB (relative gain of 3 dB) and High Boost give 1.8 dB and receive 3.2 dB (relative gain of 5 dB).

The scale factor β_R that needs to be applied to the receiver word so it is boosted 3 dB can be calculated in the following way:

$$P'_R = \sum_{t=T_{R,i}}^{T_{R,f}} (\beta_R s(t))^2 \tag{8.20}$$

$$= \beta_R^2 \sum_{t=T_{R,i}}^{T_{R,f}} s^2(t) \tag{8.21}$$

$$= \beta_R^2 P_R \tag{8.22}$$

$$\beta_R = \sqrt{\frac{P'_R}{P_R}} \tag{8.23}$$

$$= \sqrt{\frac{10^{(10\log P_R + 3)/10}}{P_R}} \tag{8.24}$$

where $T_{R,i}$ and $T_{R,f}$ define the initial and final time index that define the segment containing the receiver.

To boost a word in a sentence we first boost the word intensity and then attenuate the whole sentence. Consider first then that only the power contained in the interval

defining the receiver word is modified:

$$P_S = P_R + P_O \tag{8.25}$$

$$P'_S = P'_R + P_O \tag{8.26}$$

where P_O is the power contained in the rest of the sentence and P'_S is the sentence new power value. To find the scale factor β_S to be applied to the whole sentence to normalize its power the following:

$$P_S = \sum_{t=T_{R,i}}^{T_{R,f}} (\beta_S s'(t))^2 \tag{8.27}$$

$$= \beta_S^2 \sum_{t=1}^{T} (s'(t))^2 \tag{8.28}$$

$$= \beta_S^2 P'_S \tag{8.29}$$

$$\beta_S = \sqrt{\frac{P_S}{P'_S}} \tag{8.30}$$

$$= \sqrt{\frac{P_S}{\beta_R^2 P_R + P_O}} \tag{8.31}$$

$$= \sqrt{\frac{P_S}{\beta_R^2 P_R + P_S - P_R}} \tag{8.32}$$

$$= \sqrt{\frac{P_S}{P_R(\beta_R^2 - 1) + P_S}} \tag{8.33}$$

Instead of applying the receiver scale factor as a rectangular window we use a trapezoid window instead.

8.7.3 Listening experiment design

24 participants took part in this experiment. Each transcribed the 60 sentences once. The listening condition (two modification and unmodified) was balanced across listeners so that the whole test was covered by each group of 3 listeners, similar to previous experiment. The sentence SNR was the same as the previous experiment: -3 dB. Before the test started, the participants were given 20 sentences to transcribe to familiarize them with the task, as done in the previous experiment.

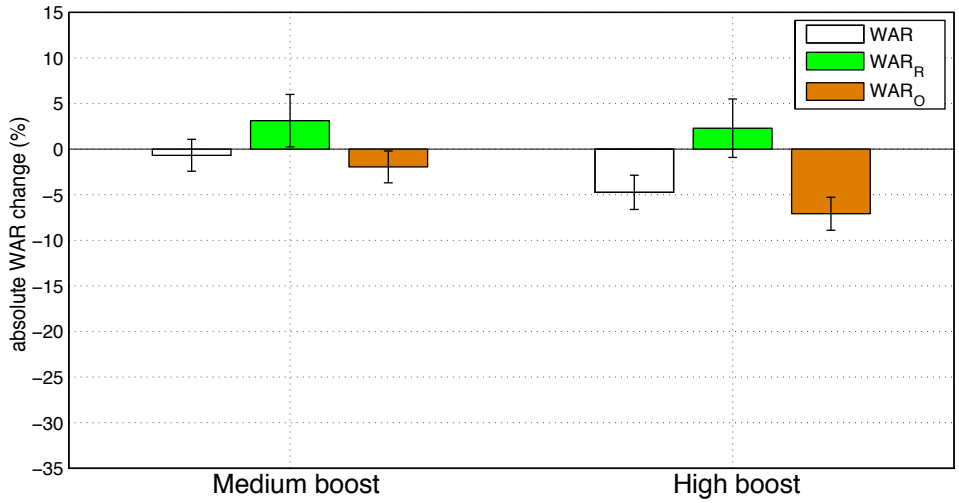


Figure 8.17: Absolute changes in WAR (in %) of Medium boost and High Boost with respect to unmodified. The WAR values obtained for unmodified speech were: WAR = 54.5 %, WAR_R = 56.0 % and WAR_O = 53.9 %.

8.7.4 Results

Fig. 8.17 shows WAR results averaged across words and listeners (deviation represents word dispersion only) in terms of absolute change compared to the unmodified case for the Medium and the High boost modifications. The acronyms WAR, WAR_R and WAR_O refer to the intelligibility of all words, receivers and others - words that were attenuated which could include words that were classified as givers or receivers depending on the sentence. As a reference, the WAR values obtained for unmodified speech were: WAR = 54.5 %, WAR_R = 56.0 % and WAR_O = 53.9 %.

We can see that for both modifications boosting a word leads to a decrease in WAR but to a smaller extent than we saw in the previous experiment. The intelligibility of givers drops for both boosting values, but by no more than a 7.0 % absolute drop, while the receivers increase up to 4.0 % word accuracy. Particularly we note that the Medium boost modification WAR_R gains are comparable to the ones obtained in the proposed modification of the previous section (around 3.0 % absolute gain) however, the loss in WAR_O is much smaller (from 29.0 % to 1.9 % absolute drop) indicating that boosting one word in a sentence is a much better strategy.

Similar to the previous experiment we will now present a more localized analysis of the differences between intelligibility scores of words in isolation and in a sentence, the effect of the receiver position and intelligibility and then revisit the overall findings.

8.7.4.1 Words in isolation versus words in sentence

As done for the previous experiment, we compare here the word accuracy results of the isolated word and the sentence experiment by finding the presentation SNR of each word in a sentence and obtaining the psychometric curves of receivers and others. These values and their standard error (which represents the variance across the words in the category) are shown in Fig. 8.18 for the Medium boost modification which had comparable WAR_R results to the proposed modification. We can see that the slope of receivers for the Medium boost modification is similar to slope for the proposed modification and that the slope for others is similar to the easy givers, as others included not only givers but also receivers.

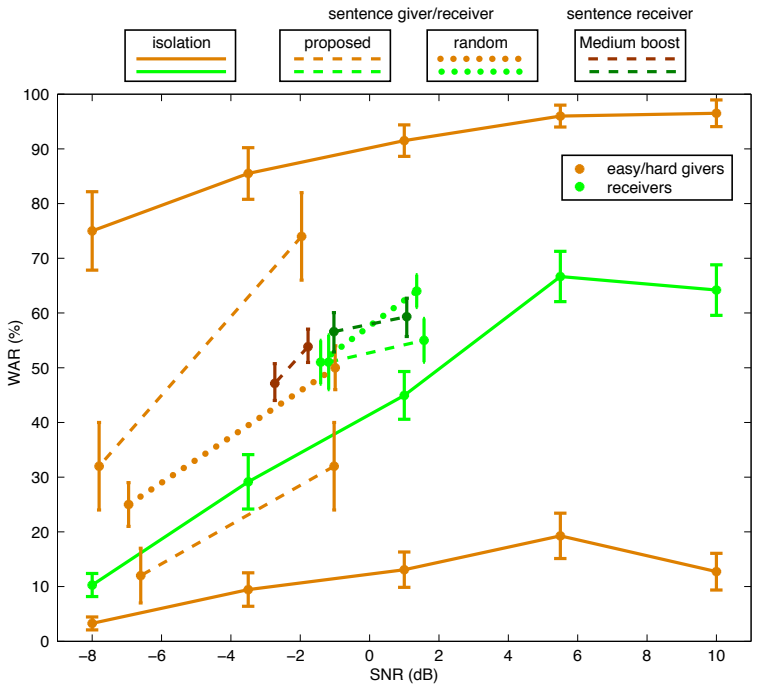


Figure 8.18: Psychometric curves of givers and receivers in isolation, in a sentence with the proposed Medium boost (darker dotted).

8.7.4.2 Receiver position and accuracy

To see the variability of the results across sentences and identify under which conditions there was an increase of word intelligibility, we present two different analyses of the effect of the receiver position and intelligibility.

Fig. 8.19 shows the WAR results, calculated for all words, receivers and others for each of the four sentence groups. The groups are named according to the position of

the receiver in the sentence: R1 (verb), R2 (first adjective), R3 (second adjective) and R4 (noun). The top figure shows the accuracy obtained for the unmodified sentence material and the figures in the middle and bottom shows the absolute changes with respect to this result obtained for the Medium and High boost modification, respectively.

It is clear from the Medium boost results that, as we found in the previous experiments, receivers gain a great deal in intelligibility when compared to the attenuation suffered to words in the others category, however boosting the last word, the noun, on average made the word less intelligible. When we increase the boost (see results for High boost) the verbs increase WAR but adjectives' intelligibility starts dropping. When the last word is boosted the other words become less intelligible at a higher rate as well.

To identify under which conditions there was an increase of word intelligibility we present in Fig. 8.20 a sentence level analysis: $\mathrm{WAR_R}$ results for each of the 60 sentences. The sentences are ordered according to the $\mathrm{WAR_R}$ obtained in the unmodified case. The continuous curve in red represents the $\mathrm{WAR_R}$ for the Medium boost (left) and the High boost (right) modifications. The dashed red curves represent these results averaged across each sentence interval. A sentence interval is taken as the range where unmodified $\mathrm{WAR_R}$ results are constant. We can clearly see that for highly intelligible words boosting can decrease $\mathrm{WAR_R}$, for both Medium and High boost modifications. It seems that if a word is more intelligible than a certain threshold then boosting is harmful and that this threshold depends on the boosting level. We can also see that the effect of the boosting value depends on the WAR of the receiver: poor receivers should be boosted more and highly intelligible receivers should not be boosted at all. Fig. 8.21 presents the same sentence ordering for the WAR results – the score for all words. We observe that boosting words selectively can increase WAR up to 15 % when the word to be boosted is a poorly intelligible one and when enough boost is applied. That is, the best strategy is to boost the most unintelligible words in the sentence and apply a RMS boost inversely proportional to the intelligibility of the word.

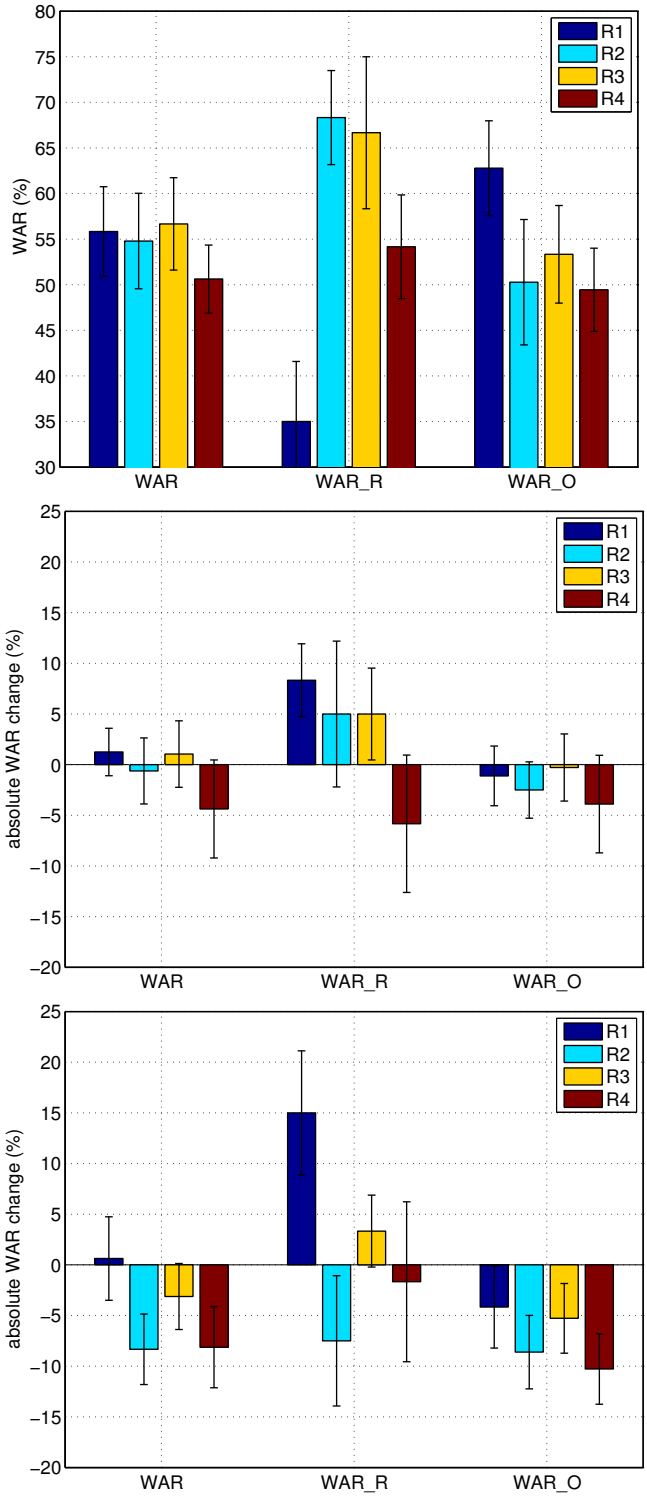
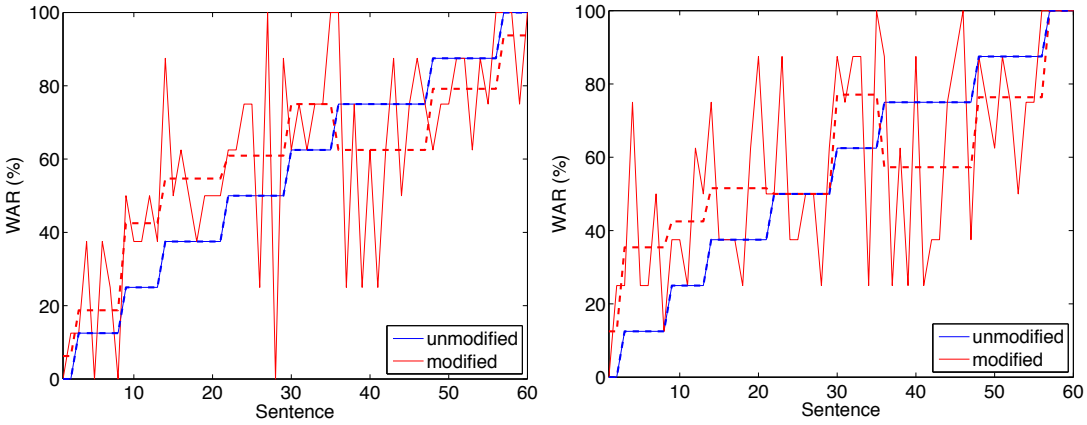
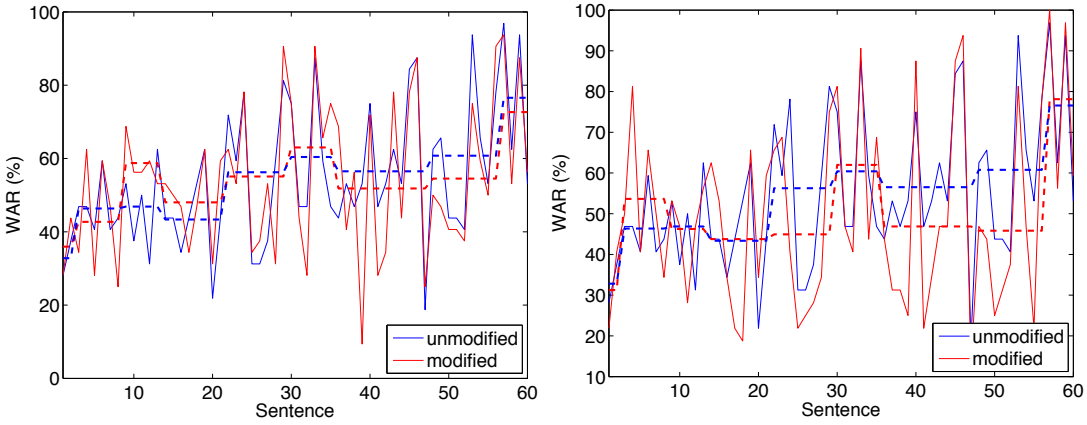


Figure 8.19: Experiment 3: Word accuracy results of unmodified (top) and word accuracy changes relative to unmodified for Medium boost (middle) and High boost (bottom). Results are calculated for each word and averaged across words for all words (WAR), receivers (WAR_R) and others (WAR_O). The numbers in the sentence group names R1, R2, R3 and R4 refer to the position of the receiver word in the sentence.



(a) Medium boost (b) High boost

Figure 8.20: Experiment 3: word accuracy of receivers (WAR_R) for each of the 60 sentences. The sentence index is ordered according to the unmodified scores (blue). Modified scores are presented on a sentence level (red continuous line) and a sentence group level (red dashed line).



(a) Medium boost (b) High boost

Figure 8.21: Experiment 3: word accuracy of all words (WAR) for each of the 60 sentences. The sentence index is ordered according to the unmodified scores (blue). Modified scores are presented on a sentence level (red continuous line) and a sentence group level (red dashed line).

8.8 Conclusion

Motivated by the fact that words are not equally confusable we explored the idea of using word-level intelligibility predictions to selectively boost words in a sentence. To do so we first evaluated the intelligibility of words in isolation through a listening experiment to find which words might benefit from SNR boosting and which words could possibly be attenuated. For this evaluation, we selected 20 words with dense and 20 words with sparse neighbourhood densities according to the OVC metric in order to cover a wide range of confusability, that is hard and easy words. The results show however that not only neighbourhood density affects intelligibility in isolation but also the speaker (a TTS voice), the masker (speech-shaped noise) and the lexical complexity of each word (duration). So instead of using ND to select words to boost – receivers – and attenuate – givers – we used the actual subjective intelligibility scores of words in isolation. Using information from that we performed two sentence experiments. We created a set of 60 sentences, each composed of four words from the isolated word experiment, two of which were a pair of giver and receiver of energy and the other two were filler words whose intensity value remained unmodified. We chose the following boost and attenuation values: 2.7dB and 6dB. The results showed that boosting a word to the detriment of another is not a good strategy, independent of the selection of the words: the intelligibility of the giver word drops an absolute value of 30% while receivers only increase 3%. Moreover, selecting the word pairs according to their intelligibility scores in isolation was not as good as choosing them randomly. Additionally, the intelligibility of words whose RMS value we did not modify also changed, showing the strategy is not appropriate. Following from these results we performed a second sentence experiment to selectively boost a word while attenuating all other words in the sentence, which also reflects a more natural way of emphasising words. Results show that we managed to decrease the intelligibility losses of givers by spreading the attenuation across all other words in the sentence: from 30% to 3% absolute drop. The gain of receivers was similar for both experiments but for a stronger boost intelligibility gains dropped. We also observed that overall intelligibility increases for those sentences where we boost a highly confusable word and that the boost value that increases intelligibility the most has to be set according to the intelligibility of this word. That is, if we had a reliable way of predicting word-level intelligibility it would be possible to increase sentence intelligibility by selectively boosting the RMS value of a highly confusable word. This is a promising result that advocates for the additional

use of word intelligibility as prior knowledge for more complex modifications. The poor word-level intelligibility prediction results using the neighbourhood density, the glimpse proportion measure and an HMM-based measure indicates that much work needs to be done in order to obtain reliable measures of word-level confusability even for the simplest scenario of words in isolation. Additionally the fact that subjective scores of words in isolation poorly predicts their scores in a sentence indicates that this prediction has to consider the context of the word in a sentence, not only for the additional linguistic cues but also for the acoustic coarticulation ones as well.

Chapter 9

Conclusions

We set out from the idea that speech perception models can be used to increase the intelligibility of synthetic speech in noise. We thought to automatically modify text-to-speech production according to the environmental noise, in much the same way as humans control their speech. It transpired that not all objective measures can reliably predict the intelligibility of synthetic speech in noise. Those that did work were based on models of the internal processing that takes place in the human auditory system. With this information our logical next step was to modify synthetic speech to improve intelligibility as defined by the scores from one of these measures. We observed in listening tests that spectral envelope modifications based on the glimpse proportion measure significantly increased intelligibility in stationary noise conditions, particularly if combined with a noise-independent strategy like dynamic range compression. To achieve similar gains in the competing speaker condition further changes to the excitation signal and duration, based on Lombard speech, were most effective. Finally, to investigate whether top-down information such as word-level intelligibility can be used as a prior to inform how much modification is required, we tested different SNR boosting strategies. We observed that selectively boosting words according to their intelligibility levels can be beneficial.

We will now present the main contributions of this work to the different areas of knowledge: objective measures of intelligibility, speech perception in noise and synthetic speech.

9.1 Contributions

9.1.1 Intelligibility prediction

We showed in this thesis that objective measures like the glimpse proportion (GP) can be used to improve intelligibility in noise. We list our main contributions to the field:

- **The development of a new spectrum-based objective measure.**

  In order to use the GP measure for cepstral coefficient manipulation we proposed an approximation, which turned the GP into a spectrum-based measure. Compared to other spectrum-based measures like the log-spectral distance, the likelihood ratio and the cepstrum and Itakura Saito distances, this reformulation of the GP displays much higher correlations with subjective intelligibility scores of synthetic speech in noise.

- **Objective measure evaluation of intelligibility enhanced speech.**

  We performed the first large scale evaluation of objective measures for the purpose of predicting the intelligibility of enhanced synthetic speech in noise. Results motivate more research into measures that are specially designed to predict intelligibility of enhanced speech in noise. The need for instance for the creation of measures that can better predict changes in speaking rate was clear. We also noted no advantage of measures that require a reference for unmodified speech (correlation and distance-based) over the GP (audibility-based). An extended version of the GP that uses a reference obtains better results showing that this additional information is useful (Tang et al., 2013).

- **Word-level intelligibility score evaluation.**

  In our final experiments, we noted that GP scores are poorly correlated with word-level subjective scores. Cooke (2009) refers to the GP and the other measures that were evaluated in this work, as macroscopic measures: they can predict an average level of intelligibility but can not give localized (word or segment) predictions. Our results illustrate the pitfalls of macroscopic measures and motivates the development of microscopic models that not only account for energetic masking at the auditory system level but at higher levels of processing as well.

9.1.2 Perception of synthetic speech in noise

During the course of this work we evaluated several intelligibility improvement strategies. The results of these experiments provide evidence about how synthetic speech is perceived in noise and how to improve it:

- **Methods to increase intelligibility of synthetic speech in noise.**

  We observed that loudness enhancement and spectral tilt flattening via spectral and temporal shaping is very effective in stationary noises or at high SNRs. For lower SNRs and for competing speaker, making prosodic changes is more beneficial.

- **Perception of synthetic speech in noise.**

  As in other studies, we found that synthetic speech is less intelligible than natural speech in noise and that the rate of deterioration is higher for increased noise levels. In particular, synthetically generated words played in isolation are very poorly recognised compared to natural speech. In a sentence, intelligibility increases because even words with many acoustically similar neighbours are more readily recognized.

- **Further investigations into noise dependency.**

  Our studies with noise-dependent and independent methods showed that in stationary noise (speech-shaped noise) dependency is not strictly necessary. It is possible to achieve comparable or better gains by simply mimicking the acoustic changes usually seen in highly intelligible natural speech with no regard paid to the noise type. For the fluctuating noise (competing speaker), noise dependency may be required. Recent work on noise-dependent duration modifications based on the GP is showing promising results (Aubanel and Cooke, 2013).

- **Contributions of fundamental frequency and speaking rate changes.**

  The use of synthetic speech made it straightforward to try perception experiments like individually changing the acoustic properties that are modified in the production of Lombard speech, that is: speaking rate, F_0 and spectral modifications (tilt and peaks). Our results indicate that changes in F_0 alone do not lead to significant changes in the intelligibility and that lowering the speaking rate is not as beneficial as changing the spectral tilt.

9.1.3 Speech synthesis

Although all contributions mentioned thus far are relevant to speech synthesis, we mention here items that contribute directly to the design of better synthetic voices:

- **A new method for cepstral extraction and cepstral modification.**

  To increase intelligibility of synthetic speech in noise we proposed new methods for cepstral extraction and modification both based on the GP measure. For the proposed method, cepstral coefficients are extracted through the maximisation of the GP and the minimisation of a distortion measure based on the Itakura Saito distance. These cepstral coefficients can then be used for training synthesis models. If noise is not stationary and is unknown at training time, the alternative is to modify cepstral coefficients generated from text at synthesis time to maximize the GP measure.

- **The identification of measures that can best predict the intelligibility of synthetic speech in noise.**

  Although there have been attempts at creating non-intrusive measures that can predict the quality of synthesizers, no proper study had been performed for the evaluation of existing intrusive measures of the intelligibility of TTS in noise. Intelligibility of synthetic speech can be quite high so the process of generating speech from text should not necessarily be measured as a source of distortion. The assumption was then made that measures created for natural speech could also be used to predict intelligibility of synthetic in noise. On performing such a study we discovered that although performance dropped it was not far below the results for natural speech and there were several measures that worked quite well even for modified synthetic speech, like for instance the GP, the Dau measure and the STOI.

- **Measuring intelligibility of synthetic speech using a measure derived from the synthesis models.**

  The intent was to use a statistical-based measure to help the prediction of word-level intelligibility combined with the neighbourhood density. Normally the lexical distance between competitors is measured using listening tests with confusion matrices built from the results. Instead we proposed to predict these distances from the synthesis models. Although the correlation scores found be-

tween the HMM-based measure and the subjective scores were low, the improvement upon using only the GP motivates further research into a combined solution that uses both the synthesis models and the audibility measure.

9.2 Future work

Intelligibility of HMM-generated synthetic speech can be very high when good quality recordings are available for training and synthetic speech is played in good listening conditions. When synthetic speech is heard in noise, recognition is compromised. Although we have showed in this thesis how one can increase recognition scores, the gap between natural plain read speech and synthetic speech built from it highlights inherent problems in the process of generating speech from text.

Perception studies with formant-based synthesizers indicate that synthetic speech is harder to perceive in adverse conditions as it lacks both the variability and redundancy found in natural speech. The same might also be said about HMM-based synthesizers. The process of vocoding (parametrization) and statistical modelling (averaging) removes important acoustic cues that are irrelevant to the decoding process in clean but become decisive in noisy conditions. The fact that we found significant differences in the intelligibility of vocoded and synthetic speech in Chapter 5 indicates that statistical modelling is responsible for removing essential perceptual cues. More investigation needs to be done to identify which aspects of training a synthetic voice are responsible for this. We believe that some of the objective measures evaluated in this work can help with this task. Additionally, they can be used as a measure for training success in a closed-loop analysis-by-synthesis strategy similar to what has been proposed in residual modelling for TTS (Maia et al., 2007) and for minimum generation error training (Wu and Wang, 2006). In order to do so one should investigate first whether the GP and the approximated version proposed in this thesis can be used to measure non-linear distortions. The GP measure as it stands can only predict the audibility of speech in noise but one could use the internal representation as proposed by the glimpsing model to measure other sort of distortions, turning the GP into a correlation or a distance-based sort of measure.

Possible extensions that can follow the work from this thesis include further analysis into the quality of enhanced synthetic speech. While speech intelligibility increases, naturalness and quality can be compromised, especially if the modified speech is heard in clean conditions. To decrease artefacts that could arise from the frame-by-frame

processing of the GP-based modification, we can imagine two solutions. One of these would be to apply the GP-based Mel cepstral modification not to the generated parameters but to the the static components of the observation vector. The maximum likelihood parameter generation (Tokuda et al., 2000) would then be responsible for smoothing the differences between consecutive frames. It is possible that updating the spectral coefficients at a slower analysis window rate would reduce artefacts while keeping similar intelligibility gains. It is to us however not clear how one should evaluate the quality of enhanced speech, whether quality scores should be given to speech listened to in noise or not, how to judge its appropriateness and what is more important in a given scenario: quality or intelligibility.

The most intelligible voice created in this work is a combination of different enhancing strategies: GP-based modification, dynamic range compression and Lombard-adapted duration and excitation. This voice obtained up to 5.8 dB of equivalent intensity gain, however it required additional recordings of Lombard speech of the speaker for which we built the voice. It would be interesting to investigate whether similar intelligibility gains could be obtained by applying cross-speaker adaptation of duration and excitation while maintaining quality and speaker similarity.

Appendix A

Objective measures of intelligibility

Here we present the results of the evaluation of objective measure evaluation without mapping the objective scores to a logistic function.

	Dau	GP	STOI	PESQ	WSS	SII	FWS	IS	CEP	LSD	LLR
r	0.90	0.86	0.87	0.80	-0.778	0.75	0.60	-0.38	-0.33	-0.33	-0.33
σ_e	0.12	0.15	0.14	0.18	0.18	0.19	0.23	0.27	0.27	0.27	0.27

Table A.1: Experiment I: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech.

	Dau	GP	STOI	PESQ	WSS	SII	FWS	IS	CEP	LSD	LLR
Case 1											
r	0.01	0.52	0.42	-0.29	0.13	0.38	0.07	-0.22	0.04	-0.14	-0.14
σ_e	0.13	0.11	0.12	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.13
Case 2											
r	0.01	0.52	0.42	-0.29	-0.51	0.38	0.07	-0.62	0.14	-0.06	-0.17
σ_e	0.13	0.11	0.12	0.13	0.11	0.12	0.13	0.11	0.13	0.13	0.13

Table A.2: Experiment I: correlation coefficient r and standard deviation of the error σ_e for *modified* synthetic speech.

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
r	0.86	0.84	0.81	-0.62	0.62	0.57	0.54	-0.49	-0.32	-0.32	-0.27
σ_e	0.10	0.12	0.13	0.17	0.17	0.18	0.18	0.19	0.21	0.21	0.21

Table A.3: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech.

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
Case 1											
r	0.71	0.73	0.61	-0.26	0.34	0.12	0.45	-0.16	-0.27	-0.25	-0.17
σ_e	0.17	0.17	0.20	0.24	0.23	0.24	0.22	0.24	0.24	0.24	0.24
Case 2											
r	0.71	0.73	0.62	0.28	0.31	0.12	0.45	0.18	0.34	0.32	0.30
σ_e	0.17	0.17	0.19	0.24	0.23	0.24	0.22	0.24	0.23	0.23	0.23

Table A.4: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *modified* synthetic speech.

	Dau	GP	STOI	WSS	PESQ	FWS	SII	IS	LSD	CEP	LLR
r	0.77	0.81	0.69	-0.58	0.27	-0.001	0.46	-0.31	-0.43	-0.39	-0.33
σ_e	0.16	0.15	0.18	0.20	0.24	0.25	0.22	0.24	0.23	0.23	0.24

Table A.5: Experiment II: correlation coefficient r and standard deviation of the error σ_e for *unmodified* synthetic speech and *LSP shift modification*.

Appendix B

Spectral gains of Lombard speech

We present here the spectral gains of the Lombard voices (natural and TTS) over the plain voices (normal and TTS) calculated at a sentence level and averaged across a set of sentences.

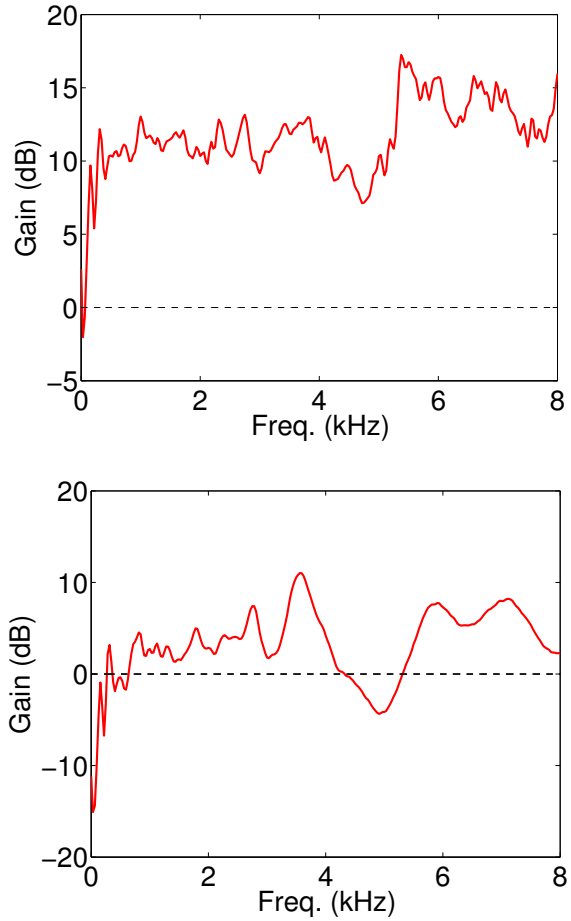


Figure B.1: Spectral gains of the natural (top) and TTS (bottom) Lombard voices.

Appendix C

Hurricane Challenge results

We present the results from all entries of the Hurricane Challenge, descriptions can be found in Cooke et al. (2013). Our entry, described in Section 7.4.1, is TTSLGP-DRC.

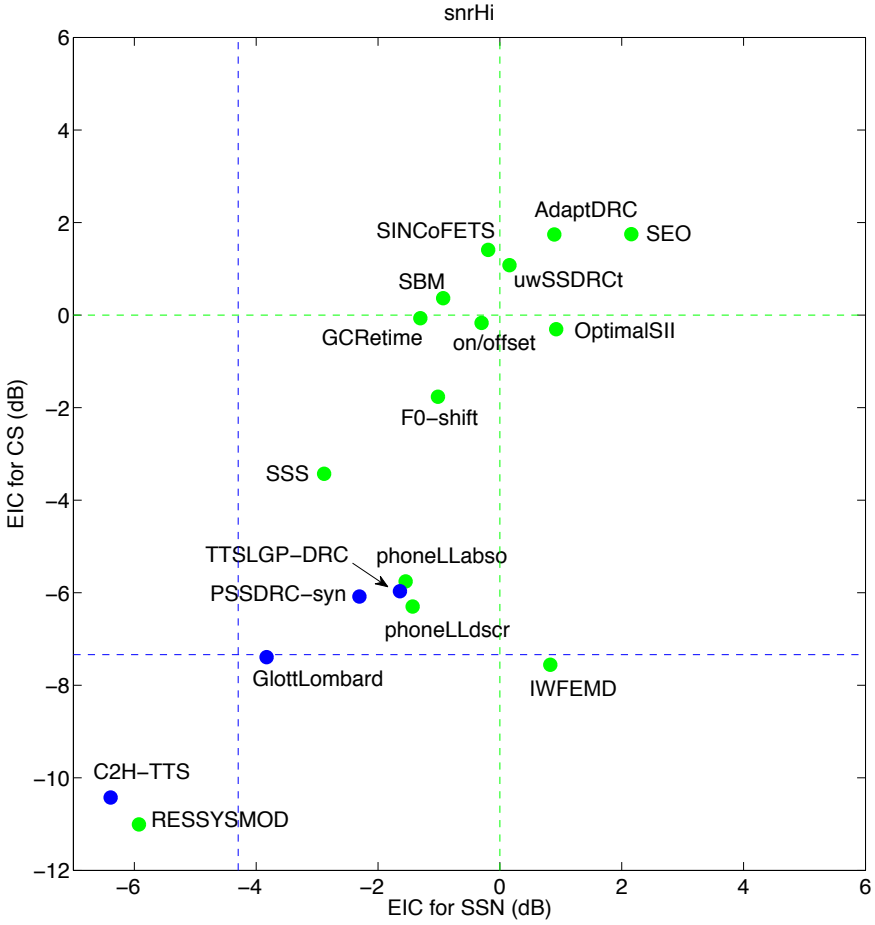


Figure C.1: Hurricane challenge results for *SNR High*: EICs in dB relative to Plain (dotted green lines) and TTS baselines (dotted blue lines) for the SSN and CS maskers. Green: natural speech entries; blue: TTS entries.

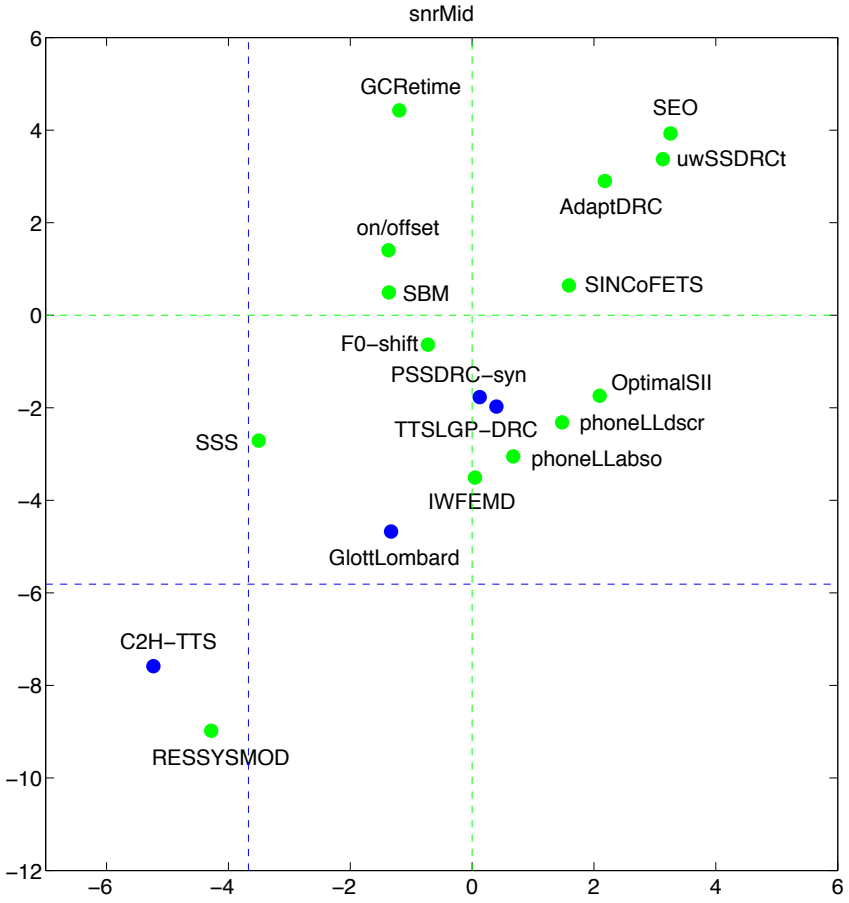


Figure C.2: Hurricane challenge results for *SNR Mid*: EICs in dB relative to Plain (dotted green lines) and TTS baselines (dotted blue lines) for the SSN and CS maskers. Green: natural speech entries; blue: TTS entries.

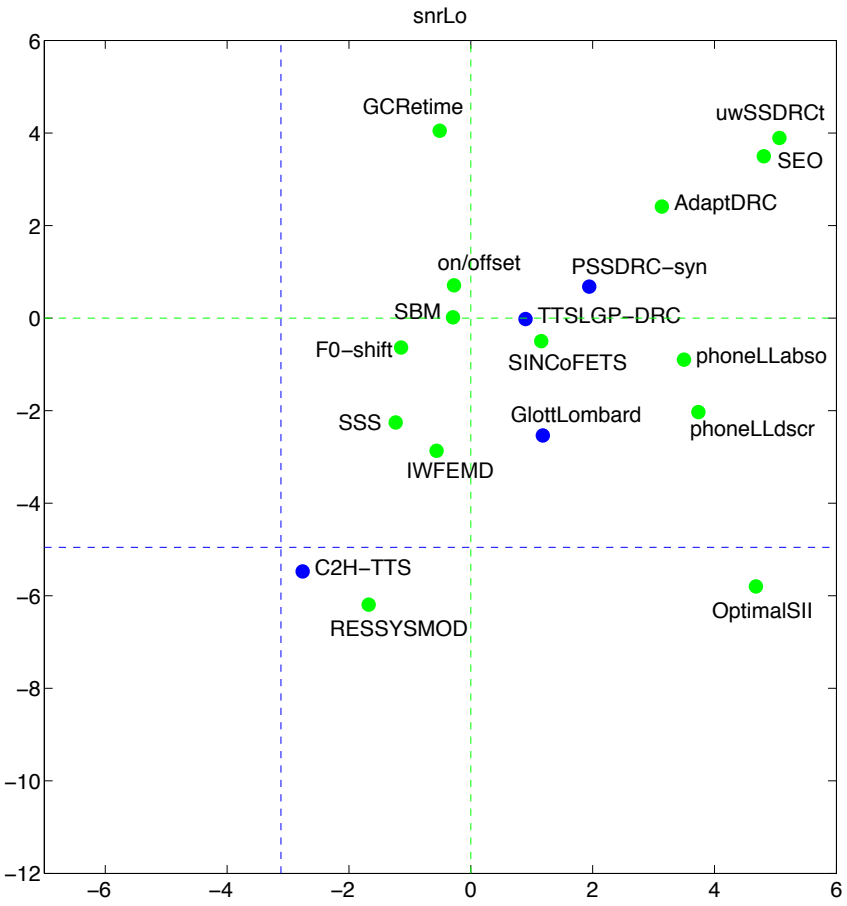


Figure C.3: Hurricane challenge results for *SNR Low*: EICs in dB relative to Plain (dotted green lines) and TTS baselines (dotted blue lines) for the SSN and CS maskers. Green: natural speech entries; blue: TTS entries.

Appendix D

Using top-down information

	Hard	Easy
Verbs	see	give
	buy	have
	sell	want
	like	change
	win	move
Adjectives	new	front
	rare	huge
	blue	next
	grey	strong
	right	strange
	main	vast
	wet	French
	spare	good
	whole	brown
	weak	close
Nouns	chair	desk
	shoe	house
	car	knife
	ball	dress
	wheel	film

Table D.1: Words classified according to their OVC ND.

	Giver (hard)	Giver (easy)	Receiver
Verbs	have	change	move
	like	see	give
	buy		sell
			want
			win
Adjectives	whole	huge	close
	right	strong	next
	weak	strange	spare
	rare	French	blue
	wet		grey
	vast		front
			good
			new
			brown
			main
Nouns	film	chair	knife
	ball	house	wheel
			dress
			car
			shoe
			desk

Table D.2: Words classified according to experiment 1 WAR.

Sentence group	Giver Receiver	Sentence
1	G1 R2	have the next French car.
	G1 R2	like the blue strange dress.
	G1 R2	buy the spare main desk.
	G1 R2	change the grey rare shoe.
	G1 R2	see the close new ball.
2	G1 R3	have the huge brown dress.
	G1 R3	like the spare wet shoe.
	G1 R3	buy the whole main knife.
	G1 R3	change the strong good car.
	G1 R3	see the blue front desk.
3	G1 R4	have the spare vast shoe.
	G1 R4	like the grey main knife.
	G1 R4	buy the huge new car.
	G1 R4	change the grey strange dress.
	G1 R4	see the spare brown wheel.
4	G2 R1	move the whole strange wheel.
	G2 R1	give the weak front chair.
	G2 R1	sell the right French film.
	G2 R1	want the strong good house.
	G2 R1	win the huge rare desk.
5	G2 R3	move the whole front knife.
	G2 R3	want the right good ball.
	G2 R3	give the weak brown desk.
	G2 R3	sell the huge main wheel.
	G2 R3	win the strong new dress.
6	G2 R4	give the strong wet knife.
	G2 R4	sell the right rare dress.
	G2 R4	want the whole vast car.
	G2 R4	win the huge brown wheel.
	G2 R4	move the weak wet shoe.

Continued on next page

Sentence group	Giver Receiver	Sentence
7	G3 R1	move the blue wet knife.
	G3 R1	give the next rare film.
	G3 R1	sell the strong French chair.
	G3 R1	want the close strange desk.
	G3 R1	win the close vast house.
8	G3 R2	change the close vast wheel.
	G3 R2	win the next strange ball.
	G3 R2	like the blue rare car.
	G3 R2	buy the spare French knife.
	G3 R2	have the grey wet shoe.
9	G3 R4	change the next wet dress.
	G3 R4	move the weak rare wheel.
	G3 R4	have the close strange car.
	G3 R4	want the right French shoe.
	G3 R4	like the next vast desk.
10	G4 R1	give the blue French ball.
	G4 R1	sell the grey good chair.
	G4 R1	want the whole good film.
	G4 R1	win the weak new chair.
	G4 R1	move the right brown house.
11	G4 R2	see the spare new film.
	G4 R2	change the blue main chair.
	G4 R2	sell the next front house.
	G4 R2	buy the close vast ball.
	G4 R2	like the grey main house.
12	G4 R3	buy the huge front film.
	G4 R3	see the whole brown chair.
	G4 R3	see the weak new house.
	G4 R3	have the right front ball.
	G4 R3	give the strong good film.

Table D.3: Sentences, proposed giver and receiver assignment.

Sentence group	Giver Receiver	Sentence
1	G1 R2	have the huge brown dress
	G1 R2	see the close new ball
	G1 R2	sell the huge main wheel
	G1 R2	sell the right French film
	G1 R2	want the whole good film
2	G1 R3	have the close strange car
	G1 R3	have the grey wet shoe
	G1 R3	buy the spare main desk
	G1 R3	buy the huge front film
	G1 R3	want the strong good house
3	G1 R4	change the blue main chair
	G1 R4	like the blue strange dress
	G1 R4	move the weak wet shoe
	G1 R4	sell the strong French chair
	G1 R4	like the spare wet shoe
4	G2 R1	like the blue rare car
	G2 R1	move the whole strange wheel
	G2 R1	sell the grey good chair
	G2 R1	give the strong good film
	G2 R1	move the whole front knife
5	G2 R3	have the next French car
	G2 R3	like the grey main knife
	G2 R3	buy the whole main knife
	G2 R3	sell the right rare dress
	G2 R3	change the next wet dress
6	G2 R4	win the strong new dress
	G2 R4	have the right front ball
	G2 R4	see the blue front desk
	G2 R4	win the huge rare desk
	G2 R4	buy the spare French knife
7	G3 R1	buy the close vast ball
	G3 R1	want the whole vast car

Continued on next page

Sentence group	Giver Receiver	Sentence
	G3 R1	move the blue wet knife
	G3 R1	change the close vast wheel
	G3 R1	see the weak new house
8	G3 R2	like the next vast desk
	G3 R2	have the spare vast shoe
	G3 R2	give the weak brown desk
	G3 R2	give the weak front chair
	G3 R2	change the grey strange dress
9	G3 R4	win the huge brown wheel
	G3 R4	like the grey main house
	G3 R4	buy the huge new car
	G3 R4	move the right brown house
	G3 R4	sell the next front house
10	G4 R1	see the whole brown chair
	G4 R1	change the strong good car
	G4 R1	see the spare brown wheel
	G4 R1	win the next strange ball
	G4 R1	change the grey rare shoe
11	G4 R2	give the next rare film
	G4 R2	want the right good ball
	G4 R2	give the blue French ball
	G4 R2	see the spare new film
	G4 R2	give the strong wet knife
12	G4 R3	move the weak rare wheel
	G4 R3	want the right French shoe
	G4 R3	win the weak new chair
	G4 R3	win the close vast house
	G4 R3	want the close strange desk

Table D.4: Sentences, random giver and receiver assignment.

Bibliography

Alku, P., Vintturi, J., and Vilkman, E. (2002). Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication*, 38(34):321 – 334.

ANSI (1997). ANSI S3.5-1997 Methods for the calculation of the speech intelligibility index.

Aubanel, V. and Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In *Proc. Interspeech*, Lyon, France.

Aubanel, V., Cooke, M., Villegas, J., and Lecumberri, M. L. G. (2011). Conversing in the presence of a competing conversation: effects on speech production. In *Proc. Interspeech*, pages 2833 – 2836, Florence, Italy.

Banos, E., Erro, D., Bonafonte, A., and Moreno, A. (2008). Flexible harmonic/stochastic modeling for HMM-based speech synthesis. In *In V Jornadas en Tecnologias del Habla*, pages 145–148, Bilbao, Spain.

Barnwell, T., III (1980). Correlation analysis of subjective and objective measures for speech quality. In *Proc. ICASSP*, volume 5, pages 706 – 709, Denver, USA.

Benoit, C. (1990). An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity. *Speech Communication*, 9(4):293–304.

Beutnagel, B., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A. (1999). The AT&T Next-Gen TTS system. In *Proc. Joint ASA, EAA and DAEA Meeting*, pages 15–19, Berlin, Germany.

Black, A. and Tokuda, K. (2005). The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. Blizzard Challenge Workshop*, Lisbon, Portugal.

Black, A. W. (2006). CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling. In *Proc. Interspeech*, pages 1762–1765, Pittsburgh, USA.

Blesser, B. (1969). Audio dynamic range compression for minimum perceived distortion. *IEEE Trans. on Audio and Electroacoustics*, 17(1):22–32.

Bouwman, G., Cranen, B., and Boves, L. (2004). Predicting word correct rate from acoustic and linguistic confusability. In *Proc. Interspeech*, Jeju Island, Korea.

Bregman, A. (1990). *Auditory scene analysis*. MIT Press, Cambridge, USA.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America.*, 120(6):4007–4018.

Buz, E. and Jaeger, T. F. (2012). Effects of phonological confusability on speech duration. In *The 25th CUNY Sentence Processing Conference*, page 46, New York, NY.

Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. (2007). Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proc. SSW*, pages 113–118, Bonn , Germany.

Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. (2008). Glottal spectral separation for parametric speech synthesis. In *Proc. Interspeech*, pages 1829–1832, Brisbane, Australia.

Cara, B. and Goswami, U. (2002). Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments and Computers*, 34:416–423.

Castellanos, A., Benedi, J., and Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, 20(1-2):23 – 35.

Cerňak, M. (2006). Unit selection speech synthesis in noise. In *Proc. ICASSP*, pages 761–764, Toulouse, France.

Chen, J.-H. and Gersho, A. (1995). Adaptive postfiltering for quality enhancement of coded speech. *IEEE Trans. on Speech and Audio Processing*, 3(1):59–71.

Christiansen, C., Pedersen, M. S., and Dau, T. (2010). Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, 52(7-8):678–692.

Cooke, M. (1993). *Modelling auditory processing and organisation*. Cambridge University Press, Cambridge, UK.

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, 31:579 – 584.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America.*, 119(3):1562–1573.

Cooke, M. (2009). Discovering consistent word confusions in noise. In *Proc. Interspeech*, pages 1887 – 1890, Brighton, UK.

Cooke, M. and Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *Journal of the Acoustical Society of America.*, 128(4):2059–2069.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: the Hurricane Challenge. In *Proc. Interspeech*, Lyon, France.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2012). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55:572–585.

Dall, R., Veaux, C., Yamagishi, J., and King, S. (2012). Analysis of speaker clustering strategies for HMM-based speech synthesis. In *Proc. Interspeech*, Portland, USA.

Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the effective signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America.*, 99(6):3615–3622.

Deller Jr., J. R., Hansen, J. H. L., and Proakis, J. G. (2000). *Discrete-Time Processing of Speech Signals*. IEEE Press Classic Reissue, Tokyo, Japan.

Di Persia, L., Milone, D., Rufiner, H. L., and Yanagida, M. (2008). Perceptual evaluation of blind source separation for robust speech recognition. *Journal of Signal Processing*, 88(10):2578 – 2583.

Di Persia, L., Yanagida, M., Rufiner, H. L., and Milone, D. (2007). Objective quality evaluation in blind source separation for speech recognition in a real room. *Journal of Signal Processing*, 87(8):1951–1965.

Digalakis, V. V., Rtischev, D., and Neumeyer, L. G. (1995). Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Trans. on Speech and Audio Processing*, 3(5):357–366.

Dreschler, W., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. *Audiology*, 40(3):148–57.

Dreschler, W. A. (2006). *Hearing in the communication society D-2-2 deliverable*. `http://hearcom.eu`.

Drugman, T. and Dutoit, T. (2010). Glottal-based analysis of the Lombard effect. In *Proc. Interspeech*, pages 2610–2613, Makuhari, Japan.

Drugman, T., Moinet, A., Dutoit, T., and Wilfart, G. (2009). Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In *Proc. ICASSP*, pages 3793–3796, Taipei, Taiwan.

Erro, D., Stylianou, Y., Navas, E., and Hernaez, I. (2012). Implementation of simple spectral techniques to enhance the intelligibility of speech using a harmonic model. In *Proc. Interspeech*, Portland, USA.

Fairbanks, G. (1958). Test of phonemic differentiation: The rhyme test. *Journal of the Acoustical Society of America.*, 30:596–600.

Falk, T., Hummel, R., and Chan, W.-Y. (2011). Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility. In *Proc. ICASSP*, pages 4480 – 4483, Prague, Czech Republic.

Falk, T. and Möller, S. (2008). Towards signal-based instrumental quality diagnosis for text-to-speech systems. *IEEE Signal Processing Letters*, 15:781–784.

Falk, T., Möller, S., Karaiskos, V., and King, S. (2008). Improving instrumental quality prediction performance for the Blizzard Challenge. In *Proc. Blizzard Challenge Workshop*, Brisbane, Australia.

Ferguson, J. D. (1980). Variable duration models for speech. In *Proc. Symposium on the Application of Hidden Markov Models to Text and Speech, IDA-CRD*, pages 143–179.

Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America.*, 88(4):1725–1736.

Fitzpatrick, M., Kim, J., and Davis, C. (2011). The effect of seeing the interlocutor on speech production in different noise types. In *Proc. Interspeech*, pages 2829 – 2832, Florence, Italy.

Flege, J. E., Bohn, O.-S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25(4):437 – 470.

Francis, A. L. and Nusbaum, H. C. (1999). The effect of lexical complexity on intelligibility. *International Journal of Speech Technology*, 3:15–25.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. ICASSP*, volume 1, pages 137–140, San Francisco, USA.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789 – 806.

Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: global effects on the utterance. In *Proc. ICSLP*, pages 2246–2249, Pittsburgh, USA.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298.

Godoy, E. and Stylianou, Y. (2012). Unsupervised acoustic analyses of normal and Lombard speech, with spectral envelope transformation to improve intelligibility. In *Proc. Interspeech*, Portland, USA.

Gomez, A. M., Schwerin, B., and Paliwal, K. (2011). Objective intelligibility prediction of speech by combining correlation and distortion based techniques. In *Proc. Interspeech*, pages 1225 – 1228, Florence, Italy.

Gramming, P., Sundberg, J., Ternstrm, S., Leanderson, R., and Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, 2:118 – 126.

Grancharov, V., Plasberg, J., Samuelsson, J., and Kleijn, W. (2008). Generalized postfilter for speech quality enhancement. *IEEE Trans. on Audio, Speech and Language Processing*, 16(1):57–64.

Gray, A., J. and Markel, J. (1976). Distance measures for speech processing. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(5):380–391.

Hall, J. L. and Flanagan, J. L. (2010). Intelligibility and listener preference of telephone speech in the presence of babble noise. *Journal of the Acoustical Society of America.*, 127(1):280–285.

Hansen, J. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, 20:151 – 173.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America.*, 115(2):833–843.

Hazan, V. and Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America.*, 130(4):2139–2152.

Hemptinne, C. (2006). *Integration of the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-Based Speech Synthesis System (HTS)*. MSc dissertation - IDIAP Research Institute, Martigny, Switzerland.

Hoover, J., Reichle, J., Van Tasell, D., and Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax. *Journal of Speech and Hearing Research*, 30(3):425–431.

Howell, P., Barry, W., and Vinson, D. (2006). Strength of British English accents in altered listening conditions. *Perception and Psychophysics*, 68:139–153.

Hu, Y. and Loizou, P. (2007a). A comparative intelligibility study of speech enhancement algorithms. In *Proc. ICASSP*, volume 4, pages 561–564, Honolulu, USA.

Hu, Y. and Loizou, P. (2008a). Evaluation of objective quality measures for speech enhancement. *IEEE Trans. on Audio, Speech and Language Processing*, 16(1):229–238.

Hu, Y. and Loizou, P. C. (2006). Evaluation of objective measures for speech enhancement. In *Proc. Interspeech*, pages 1447 – 1450, Pittsburgh, USA.

Hu, Y. and Loizou, P. C. (2007b). A comparative intelligibility study of single-microphone noise reduction algorithms. *Journal of the Acoustical Society of America.*, 122(3):1777–1786.

Hu, Y. and Loizou, P. C. (2008b). A new sound coding strategy for suppressing noise in cochlear implants. *Journal of the Acoustical Society of America.*, 124(1):498–509.

Huang, D.-Y., Rahardja, S., and Ong, E. P. (2010). Lombard effect mimicking. In *Proc. SSW*, pages 258–263, Kyoto, Japan.

Hunt, A. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. ICASSP*, pages 373–376, Atlanta, USA.

IEEE (1969). IEEE recommended practice for speech quality measurement. *IEEE Trans. on Audio and Electroacoustics*, 17(3):225 – 246.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. ICASSP*, volume 8, pages 93–96, Boston, USA.

Imai, S. and Furuichi, C. (1988). Unbiased estimator of log spectrum and its application to speech signal processing. In *Proc. EURASIP*, pages 203–206, Grenoble, France.

ISO 532 (1975). Acoustics - method for calculating loudness level.

Itakura, F. (1975a). Line spectrum representation of linear predictor coefficients of speech signals. *Journal of the Acoustical Society of America.*, 57(S1):S35–S35.

Itakura, F. (1975b). Minimum prediction residual principle applied to speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 23(1):67–72.

Itakura, F. and Saito, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electr. and Commun. in Japan*, 52-A:36–43.

ITU (2004). ITU-T Rec. P.563 Single ended method for objective speech quality assessment in narrowband telephony applications.

Jepsen, M., Ewert, S., and Dau, T. (2008). A computational model of human auditory signal processing and perception. *Journal of the Acoustical Society of America.*, 124(1):422–438.

Jokinen, E., Takanen, M., Vainio, M., and Alku, P. (2013). An adaptive post-filtering method producing an artificial lombard-like effect for intelligibility enhancement of narrowband telephone speech. *Computer Speech and Language*, (0):–.

Jokinen, E., Yrttiaho, S., Pulakka, H., Vainio, M., and Alku, P. (2012). Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech (in press). *Journal of the Acoustical Society of America.*, 132(6):3990–4001.

Juang, B. H. and Rabiner, L. R. (1985). A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64:391–408.

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America.*, 93(1):510–524.

Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *Journal of the Acoustical Society of America.*, 117(4):2224–2237.

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Proc. ICASSP*, volume 2, pages 1303–1306, Munich, Germany.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. MAVEBA*, Florence, Italy.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207.

Kim, H. and Lee, H. (2000). Spectral peak-weighted liftering of cepstral coefficients for speech recognition. *IEICE Trans. Inf. Syst.*, 83(7):1540–1549.

King, S. and Karaiskos, V. (2010). The Blizzard Challenge 2010. In *Proc. Blizzard Challenge Workshop*, Kyoto, Japan.

Kjems, U., Boldt, J., Pedersen, M., Lunner, T., and Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *Journal of the Acoustical Society of America.*, 126(3):1415–1426.

Klatt, D. (1982). Prediction of perceived phonetic distance from critical-band spectra:a first step. In *Proc. ICASSP*, volume 7, pages 1278–1281, Paris, France.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America.*, 67:971–995.

Koishida, K. (1998). *Low bit rate speech coding based on Mel-generalized cepstral analysis*. Tokyo Institute of Technology, Tokyo, Japan.

Koishida, K., Tokuda, K., Kobayashi, T., and Imai, S. (1996). CELP coding system based on mel-generalized cepstral analysis. In *Proc. ICSLP*, volume 1, pages 318 –321, Philadelphia, USA.

Koishida, K., Tokuda, K., Kobayashi, T., and Imai, S. (2000). Spectral representation of speech based on mel-generalized cepstral coefficients and its properties. *Electronics and Communications in Japan*, 83(3):50–59.

Kokkinakis, K. and Loizou, P. C. (2011). Evaluation of objective measures for quality assessment of reverberant speech. In *Proc. ICASSP*, pages 2420 –2423, Prague, Czech Republic.

Koul, R. K. and Allen, G. D. (1993). Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in presence of noise. *Journal of Speech and Hearing Research*, 36(4):790–798.

Kubichek, R., Atkinson, D., and Webster, A. (1991). Advances in objective voice quality assessment. In *Glob. Telecomm. Conf.*, volume 3, pages 1765–1770, Phoenix, USA.

Langner, B. and Black, A. W. (2005). Improving the understandability of speech synthesis by modeling speech in noise. In *Proc. ICASSP*, volume 1, pages 265–268, Philadelphia, USA.

Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171 – 185.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W. J. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers.

Ling, Z., Wu, Y., Wang, Y., Qin, L., and Wang, R. (2006). USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Proc. Blizzard Challenge Workshop*, Pittsburgh, USA.

Ling, Z.-H. and Dai, L.-R. (2012). Minimum Kullback-Leibler divergence parameter generation for HMM-based speech synthesis. *IEEE Trans. on Audio, Speech and Language Processing*, 20(5):1492–1502.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America.*, 86(2):566–581.

Loizou, P. and Kim, G. (2011). Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE Trans. on Audio, Speech and Language Processing*, 19(1):47–56.

Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. CRC Press, Boca Raton, USA, 1 edition.

Lombard, E. (1911). Le signe d'élévation de la voix [the sign of the elevation of the voice]. *Annales des maladies de l'oreille et du larynx*, 37:101–119.

Lu, Y. and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *Journal of the Acoustical Society of America.*, 124(5):3261–3275.

Lu, Y. and Cooke, M. (2009a). Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *Journal of the Acoustical Society of America.*, 126(3):1495–1499.

Lu, Y. and Cooke, M. (2009b). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262.

Luce, P. and Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1–36.

Luce, P. A. and McLennan, C. T. (2008). *Spoken Word Recognition: The Challenge of Variation*, pages 591–609. Blackwell Publishing Ltd, Oxford, UK.

Ma, J., Hu, Y., and Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *Journal of the Acoustical Society of America.*, 125(5):3387–3405.

Ma, J. and Loizou, P. C. (2011). SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. *Speech Communication*, 53(3):340–354.

Macleod, A. and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2):131–141.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, 31(1):133–156.

Maia, R., Akamine, M., and Gales, M. J. (2013). Complex cepstrum for statistical parametric speech synthesis. *Speech Communication*, 55(5):606 – 618.

Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. (2007). An excitation model for HMM-based speech synthesis based on residual modeling. In *Proc. Speech Synthesis Workshop*, pages 131–136, Bonn , Germany.

McLoughlin, I. and Chance, R. (1997). LSP-based speech modification for intelligibility enhancement. In *Proc. Digital Signal Processing*, volume 2, pages 591–594, Santorini , Greece.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In *The Oxford handbook of psycholinguistics*, pages 37–53.

Mirenda, P. and Beukelman, D. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*, 3(3):120–128.

Möller, S. and Falk, T. H. (2009). Quality prediction for synthesized speech: Comparison of approaches. In *Conf. on Acoustics*, pages 1168–1171, Rotterdam, The Netherlands.

Moore, B. C. J. and Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica*, 82:335–345.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.

Nicolao, M., Latorre, J., and Moore, R. K. (2012). C2H A computational model of H&H-based phonetic contrast in synthetic speech. In *Proc. Interspeech*, Portland, USA.

Niederjohn, R. J. and Grotelueschen, J. H. (1976). The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(4):277–282.

Norrenbrock, C., Hinterleitner, F., Heute, U., and Möller, S. (2012). Towards perceptual quality modeling of synthesized audiobooks. In *Proc. Blizzard Challenge Workshop*, Portland, USA.

Nye, P. and Gaitenby, J. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). *Haskins Laboratories Status Report on Speech Research*, 33:77–91.

Patel, R., Everett, M., and Sadikov, E. (2006). Loudmouth: Modifying text-to-speech synthesis in noise. In *ACM SIGACCESS Conf. on Computers and Accessibility*, pages 227–228, New York, USA.

Patel, R. and Schell, K. W. (2008). The influence of linguistic content on the Lombard effect. *Journal of Speech, Language and Hearing Research*, 51:209–220.

Petkov, P., Kleijn, B., and Henter, G. (2012). Enhancing subjective speech intelligibility using a statistical model of speech. In *Proc. Interspeech*, Portland, USA.

Picart, B., Drugman, T., and Dutoit, T. (2011). Continuous control of the degree of articulation in HMM based speech synthesis. In *Proc. Interspeech*, pages 1797 – 1800, Florence, Italy.

Picart, B., Drugman, T., and Dutoit, T. (2013). Analysis and HMM-based synthesis of hypo and hyperarticulated speech (in press). *Computer Speech and Language*.

Picheny, M., Durlach, N., and Braida, L. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, 28:96–103.

Pisoni, D., Nusbaum, H., and Greene, B. (1985). Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73(11):1665–1676.

Pitz, M. and Ney, H. (2005). Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. on Speech and Audio Processing*, 13(5):930–944.

Quackenbush, S., Barnwell, T., and Clements, M. (1988). *Objective measures of speech quality*. Prentice Hall, New Jersey, USA.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. (2008). HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In *Proc. Interspeech*, pages 1881–1884, Brisbane, Australia.

Raitio, T., Suni, A., Vainio, M., and Alku, P. (2011a). Analysis of HMM-based Lombard speech synthesis. In *Proc. Interspeech*, pages 2781 – 2784, Florence, Italy.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. (2011b). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans. on Audio, Speech and Language Processing*, 19(1):153–165.

Raitio, T., Takanen, M., Santala, O., Suni, A., Vainio, M., and Alku, P. (2012). On measuring the intelligibility of synthetic speech in noise – Do we need a realistic noise environment? In *Proc. ICASSP*, pages 4025–4028, Kyoto, Japan.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *Journal of the Acoustical Society of America.*, 120(6):3988–3997.

Richmond, K., Clark, R., and Fitt, S. (2010). On generating Combilex pronunciations via morphological analysis. In *Proc. Interspeech*, pages 1974–1977, Makuhari, Japan.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. ICASSP*, volume 2, pages 749–752, Salt Lake City, USA.

Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K. (1992). ATR ν-TALK speech synthesis system. In *Proc. ICSLP*, pages 483–486, Banff, Canada.

Saheer, L., Garner, P. N., Dines, J., and Liang, H. (2010). VTLN adaptation for statistical speech synthesis. In *Proc. ICASSP*, pages 4838–4841, Dallas, USA. IEEE.

Saheer, L., Yamagishi, J., Garner, P., and Dines, J. (2012). Combining vocal tract length normalization with hierarchial linear transformations. In *Proc. ICASSP*, pages 4493 –4496, Kyoto, Japan.

Sauert, B. and Vary, P. (2006). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proc. ICASSP*, pages 493–496, Toulouse, France.

Sauert, B. and Vary, P. (2010). Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In *Proc. ITG-Fachtagung Sprachkommunikation*, volume 9, Bochum, Germany.

Sauert, B. and Vary, P. (2011). Near end listening enhancement considering thermal limit of mobile phone loudspeakers. In *Proc. Conf. on Elektronische Sprachsignalverarbeitung*, volume 61, pages 333–340, Aachen, Germany.

Sauert, B. and Vary, P. (2012). Near-end listening enhancement in the presence of bandpass noises. In *Proc. of ITG-Fachtagung Sprachkommunikation*, volume 10, pages 195–198, Berlin, Germany.

Scarborough, R. (2010). Lexical and contextual predictability: Confluent effects on the production of vowels. *Papers in Laboratory Phonology X*, pages 557–586.

Schroeder, M. R. and Atal, B. (1985). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *Proc. ICASSP*, volume 10, pages 937–940, Florida, USA.

Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *Journal Acoust. Soc. Jpn.(E)*, 21(2):79–86.

Skowronski, M. D. and Harris, J. G. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558.

Stent, A., Syrdal, A., and Mishra, T. (2011). On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *Proc. on Computers and Accessibility*, pages 211–218, Dundee, UK.

Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. (1988). Effects of noise on speech production: Acoustic and perceptual analysis. *Journal of the Acoustical Society of America.*, 84:917–928.

Suni, A., Raitio, T., Vainio, M., and Alku, P. (2010). The GlottHMM speech synthesis entry for Blizzard Challenge 2010. In *Proc. Blizzard Challenge Workshop*, Kyoto, Japan.

Syrdal, A. K., Bunnell, H. T., Hertz, S. R., Mishra, T., Spiegel, M. F., Bickley, C., Rekart, D., and Makashay, M. J. (2012). Text-To-Speech intelligibility across speech rates. In *Proc. Interspeech*, Portland, USA.

Taal, C., Hendriks, R., Heusdens, R., Jensen, J., and Kjems, U. (2009). An evaluation of objective quality measures for speech intelligibility prediction. In *Proc. Interspeech*, pages 1947–1950, Brighton, UK.

Taal, C. H., Hendriks, R. C., and Heusdens, R. (2012). A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. In *Proc. ICASSP*, pages 4061–4064, Kyoto, Japan.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. ICASSP*, pages 4214–4217, Dallas, USA.

Talkin, D. (1995). *A robust algorithm for pitch tracking (RAPT)*, pages 495–518. Elsevier, New York, USA.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc. ICASSP*, pages 805–808, Salt Lake City, USA.

Tang, Y. and Cooke, M. (2010). Energy reallocation strategies for speech enhancement in known noise conditions. In *Proc. Interspeech*, pages 1636–1639, Makuhari, Japan.

Tang, Y. and Cooke, M. (2011). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Proc. Interspeech*, pages 345–348, Florence, Italy.

Tang, Y. and Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Proc. Interspeech*, Portland, USA.

Tang, Y., Cooke, M., and Valentini-Botinhao, C. (2013). A distortion-weighted glimpse-based intelligibility metric for modied and synthetic speech. In *Proc. SPIN*, page 32, Vitoria, Spain.

Toda, T. and Tokuda, K. (2005). Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Proc. Interspeech*, pages 2801–2804, Lisbon, Portugal.

Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. Syst.*, E90-D(5):816–824.

Tokuda, K., Kobayashi, T., and Imai, S. (1995). Adaptive cepstral analysis of speech. *IEEE Trans. on Speech and Audio Processing*, SA-3(6):481–489.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis — a unified approach to speech spectral estimation. In *Proc. ICSLP*, volume 3, pages 1043–1046, Yokohama, Japan.

Tokuda, K., Kobayashi, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis. In *Proc. ICASSP*, pages 1315–1318, Istanbul, Turkey.

Tokuda, K., Kobayashi, T., Yamamoto, R., and Imai, S. (1989). Spectral estimation of speech based on generalized cepstral representation. *Trans. (A) I.E.I.C.E.*, J72-A:457–465.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (2002). Multi-space probability distribution HMM. *IEICE Trans. Inf. Syst.*, E85-D(3):455–464.

Tokuda, K., Zen, H., Yamagishi, J., Black, A., Masuko, T., and Sako, S. (2009). *The HMM-based speech synthesis system (HTS) version 2.1.* `http://hts.sp.nitech.ac.jp/`.

Tribolet, J., Noll, P., McDermott, B., and Crochiere, R. (1978). A study of complexity and quality of speech waveform coders. In *Proc. ICASSP*, volume 3, pages 586–590, Oklahoma, USA.

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak E-NG-L-I-SH? a comparison of foreigner- and infant-directed speech. *Speech Communication*, 49:2–7.

Valentini-Botinhao, C., Godoy, E., Stylianou, Y., Sauert, B., King, S., and Yamagishi, J. (2013a). Improving intelligibility in noise of HMM-generated speech via noise-dependent and -independent methods. In *Proc. ICASSP*, pages 7854–7858, Vancouver, Canada.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., and Zen, H. (2012a). Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. In *Proc. ICASSP*, pages 3997–4000, Kyoto, Japan.

Valentini-Botinhao, C., Wester, M., Yamagishi, J., and King, S. (2013b). Using neighbourhood density and selective snr boosting to increase the intelligibility of synthetic speech in noise. In *Proc. Speech Synthesis Workshop*, Barcelona, Spain.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011a). Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In *Proc. Interspeech*, pages 1837 – 1840, Florence, Italy.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011b). Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise. In *Proc. ICASSP*, pages 5112–5114, Prague, Czech Republic.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012b). Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise. In *Proc. SAPA*, Portland, USA.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012c). Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise. In *Proc. Interspeech*, Portland, USA.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012d). Using an intelligibility measure to create noise robust cepstral coefficients for HMM-based speech synthesis. In *Proc. LISTA Workshop*, page 86, Edinburgh, UK.

Valentini-Botinhao, C., Yamagishi, J., King, S., and Maia, R. (2013c). Intelligibility enhancement of HMM-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion. *Computer Speech and Language (in press)*.

Valentini-Botinhao, C., Yamagishi, J., King, S., and Stylianou, Y. (2013d). Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise. In *Proc. Interspeech*, Lyon, France.

van Wijngaarden, S. J., Steeneken, H. J., and Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *Journal of the Acoustical Society of America.*, 111:1906–1916.

Venkatagiri, H. S. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America.*, 113(4):2095–2104.

Villegas, J., Cooke, M., and Mayo, C. (2012). The role of durational changes in the Lombard speech advantage. In *Proc. LISTA Workshop*, page 87, Edinburgh, UK.

Winters, S. and Pisoni, D. (2003). Perception and comprehension of synthetic speech. *Research on Spoken Language Processing – Progress Report No. 26*, pages 95–138.

Winters, S. J. and Pisoni, D. B. (2006). Speech synthesis, perception and comprehension of. In Brown, K., editor, *Encyclopedia of Language & Linguistics (Second Edition)*, pages 31 – 49. Elsevier, Oxford, UK, second edition.

Wolters, M. K., Isaac, K. B., and Renals, S. (2010). Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proc. Speech Synthesis Workshop*, pages 136–141, Tokyo, Japan.

Womack, B. and Hansen, J. (1996). Classification of speech under stress using target driven features. *Speech Communication*, 20:131 – 150.

Wu, Y. and Tokuda, K. (2008). Minimum generation error training with direct log spectral distortion on LSPs for HMM-based speech synthesis. In *Proc. Interspeech*, pages 577–580, Brisbane, Australia.

Wu, Y.-J. and Tokuda, K. (2009). Minimum generation error training by using original spectrum as reference for log spectral distortion measure. In *Proc. ICASSP*, pages 4013–4016, Taipei, Taiwan.

Wu, Y.-J. and Wang, R.-H. (2006). Minimum generation error training for HMM-Based speech synthesis. In *Proc. ICASSP*, volume 1, pages I89–I92, Toulouse, France.

Yamagishi, J. (2006). *Average-Voice-Based Speech Synthesis*. Tokyo Institute of Technology, Tokyo, Japan.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Trans. on Audio, Speech and Language Processing*, 17(1):66 –83.

Yamagishi, J., Ling, Z., and King, S. (2008a). Robustness of HMM-based speech synthesis. In *Proc. Interspeech*, pages 581–584, Brisbane, Australia.

Yamagishi, J., Masuko, T., and Kobayashi, T. (2004). HMM-based expressive speech synthesis-Towards TTS with arbitrary speaking styles and emotions. In *Proc. of Special Workshop in Maui, SWIM*, volume 34, Maui, USA.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., and Tokuda, K. (2008b). Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge. In *Proc. Blizzard Challenge Workshop*, volume 5, Brisbane, Australia.

Yamato, O., Tomoki, T., Hiroshi, S., and Shikano, K. (2006). Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proc. Interspeech*, pages 2266 – 2269, Pittsburgh, USA.

Yao, Y. (2011). *The effects of phonological neighborhoods on pronunciation variation in conversational speech*. PhD thesis - University of California, California, USA.

Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K., and Shaiman, S. (2007). Speech signal modification to increase intelligibility in noisy environments. *Journal of the Acoustical Society of America.*, 122(2):1138–1149.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. In *Proc. ICSLP*, pages 29–32, Sydney, Australia.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2347–2350, Budapest, Hungary.

Yoshimura, T., Tokuda, K., Masukom, T., Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis. In *Proc. Eurospeech*, pages 2263–2267, Aalborg, Denmark.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.-Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The Hidden Markov Model Toolkit (HTK) version 3.4*. `http://htk.eng.cam.ac.uk/`.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. Syst.*, E90-D(1):325–333.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zen, H., Tokuda, K., and Kitamura, T. (2007b). Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, 21(1):153–173.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007c). A hidden semi-markov model-based speech synthesis system. *IEICE - Trans. Inf. Syst.*, E90-D(5):825–834.

Zhao, Y. and Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics*, 37(2):231 – 247.

Zorilă, T. C., Kandia, V., and Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Proc. Interspeech*, Portland, USA.